Railroads, Economic Development, and the Demographic Transition in the United States

Ori Katz[1]

August 2018

Abstract

This paper estimates the impact of railroads in the United States between 1850 and 1910 on economic development, fertility, and human capital. A novel identification strategy, which relies on a dynamic instrument, allows me to control for unobservables using county fixed effects. I find that railroads shifted the distribution of occupations and industries, had a large positive effect on human capital levels, and a large negative effect on fertility rates. Further analysis suggests that the impact of railroads was larger in counties that were initially more developed. I examine possible mechanisms that drive the effects and lead to this heterogeneity.

[1] Brown University, Economics Department; email: ori_katz@brown.edu
I am grateful to Moshe Hazan, Omer Moav, David Weil, Oded Galor, Raphael Franck and participants of several seminars in Tel Aviv University and Brown University for helpful comments and suggestions.
This research was supported by the Israel Science Foundation (grant No.59/1/)

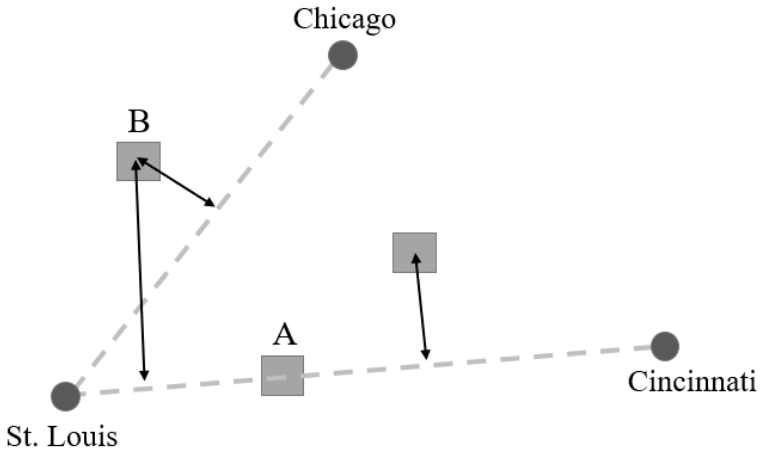
1. Introduction

This study provides evidence for a causal effect of railroads on economic development, fertility, and human capital, using panel data of 1,490 US counties for the period 1850-1910. Estimating the magnitude of these effects is important because railroads were the dominant form of freight transportation during this period, and because the decline in fertility rates and the increasing returns to human capital played a critical part in the transition from the Malthusian stagnation to the modern regime of constant economic growth (Galor, 2011).

The rapid expansion of the railroads in the US during the second half of the 19th century connected remote counties to the national trade network, enabling us to identify the effects of the induced economic development on human capital fertility, effects which are usually gradual and harder to see. However, both the timing and the location of railroad construction might have been endogenous. Reverse causality and unobserved variables do not allow us to estimate the effect of railroads directly, using a simple OLS approach. Therefore, in order to identify a causal relationship, I use the growth of new major cities as a natural experiment.

An example is shown in Figure 1. St. Louis, Cincinnati, and Chicago are three major cities that experienced rapid population growth during the second half of the 19th century. While the cities got more developed, large investments were made in transportation infrastructures that connected them to each other and to other major cities. The exact routes of the transportation infrastructures might have been endogenous, but due to cost considerations their routes resembled straight connecting lines between the cities. County A in Figure 1, which happened to be located between St. Louis and Cincinnati, got access to new transportation infrastructures because of its location, and experienced exogenous economic development, which was not related to attributes of the local geography or population.

Figure 1: An Example for the Identification Strategy



Using distance to the straight connecting lines as an instrument for the distance to actual railroads allows me to capture the exogenous effect of the new transportation infrastructures on economic development, fertility, and human capital.

Similar identification strategies were used to identify the effect of railroads by Atack, Haines, and Margo (2008), Atack, Bateman, Haines, and Margo (2010), and Banerjee Duflo and Qian (2012). A major difference between those studies and this one is the dynamic dimension of my instrument: the distance to connecting lines between major cities changed with time because of the appearance of new major cities. For example, in 1850 Chicago was only the 21st largest city in the US, while in 1910 it was the second largest city. For county B in Figure 1, the distance to the nearest connecting line in 1850 is relatively large, but not during later years, when Chicago was also considered a major city. The dynamic nature of this natural experiment allows me to control for unobservables using county fixed effects (as well as year fixed effects that control for time trends). To mitigate concerns regarding the endogenous location of the major cities, I also control the distance to the nearest major city in each period.

Using several historical data sources, I construct a 7-period panel data for all the counties that existed between 1850 and 1910. Most of the results are limited to 1,490 counties east of the 95° line of longitude, because the western counties were not highly populated at the time, their borders changed considerably, and the empirical strategy makes less sense for counties that were far away from the largest cities (see Figure 2).

The explanatory variable we are interested in is the distance from the centroid of a county to the nearest railroad, which is instrumented by the distance to the nearest straight connecting line. The outcomes include variables for fertility, human capital, and economic development. Two measures are used for fertility: the number of children aged 5-18 per woman aged 20-44 (survival fertility), and the total fertility rate of women aged 15-44. The advantage of the first measure is its availability for all the periods, and its insensitivity to the trends in infant mortality, which might have affected birth decisions made by parents. The advantage of the second measure is that it is more closely related to what economists and demographers usually mean when they talk about fertility, but it is only available for some of the periods. Human capital is measured by the literacy rates of adult males, and by an occupational socioeconomic score (based on Dunkan, 1961), which also captures some aspects of economic development. Other aspects of economic development are captured by the share of non-agricultural male workers, and by the value of manufacturing output per capita.

A descriptive analysis of the outcome variables establishes a significant positive correlation between the economic development variables and literacy rates, and a significant negative correlation between the economic development variables and the fertility variables. A descriptive analysis of the effect of railroads, without using the instrument, establishes a significant positive correlation between the distance to the nearest railroad and the fertility variables, and significant negative correlations between the distance to the nearest railroad and the economic development and human capital variables. These correlations hold

also after controlling county and year fixed effects. Furthermore, it seems that even without using our natural experiment there are no trends in most of the outcome variables prior to the arrival of the railroads, while clear trends emerge after the arrival of the railroads.

To justify the use of the instrument I show that there is a strong correlation between the distance to the connecting lines and the distance to actual railroads, after controlling fixed effects for years and counties and the distance to the nearest major city. Furthermore, I show that prior to the emergence of the new major cities, counties along the future connecting lines were no more developed than other counties.

The main results establish a significant causal effect of the distance to railroads, instrumented by the distance to connecting lines, on economic development, fertility, and human capital. Reducing the distance to the nearest railroad by 10% increases the occupational socioeconomic score by 1.17%, increases the share of non-agriculture male workers by 3.24%, increases the value of manufacturing output per capita by 2.27%, decreases survival fertility by 1.75%, decreases total fertility rate by 2.55%, and increases the share of literate adult males by 1.17%. Compared with the distribution of the outcome variables during the period, those elasticities represent a large effect on fertility and literacy, a more moderate effect on the occupation and industry structure, and a small effect on the development of the manufacturing sector. Those results are robust for controlling for the distance to waterways, the sex ratio, the share of foreign immigrants and the share of white population, as well as for different specifications of the instrument and the sample group.

Heterogeneity analysis suggests that the effects were relatively larger in counties that were more developed in 1850 and relatively smaller in the less developed counties, due to specialization in skilled-intensive industries in the more developed counties. This result confirms the prediction of Galor and Mountford (2008) about asymmetric gains from trade due to differences in

specialization, a mechanism that might explain the increasing gaps between industrial societies and other societies since the 19th century (the "Great Divergence"). Further analysis suggests that the economic development induced by railroads was accompanied by an increase in the age of marriage and in the share of foreign immigrants. However, it was not accompanied by a change in the sex ratio that might decrease fertility in a "mechanical" way.

This study furthers our knowledge in three important ways. First, previous studies of the effect of railroads or industrialization on economic development, fertility, and human capital were based on cross-sectional data, could not control county-level fixed effects, and thus might be biased because of unobserved variables, while this study provides an identification strategy that controls county-level fixed effects using panel data and a dynamic instrument. Second, the study creates a link between the railroad literature and the long-term growth literature, by analyzing the effects of railroads on the Demographic Transition and on human capital accumulation - the basic ingredients for long-term growth. My results might explain the long-term persistence in the effect of transportation infrastructure found in other studies. And third, as mentioned before, the rich data sources used in this study allow for a heterogeneity analysis of the effect of railroads, which provides evidence for an important mechanism related to the Great Divergence.

The paper is organized as follows. The next section surveys the relevant theoretical and empirical literature. Section 3 presents the data and a descriptive analysis of the main variables and the relationships between them. Section 4 discusses the empirical strategy and the validation of the instrument. Section 5 presents the effect of railroads on economic development, fertility, and human capital. Section 6 examines heterogeneity in the effect of railroads. Section 7 discusses some of the mechanisms that might drive the effect. Section 8 concludes.

2. Related Literature

Four different strands of the economic literature are relevant for this paper. The first one includes theoretical studies of the mechanisms behind the Demographic Transition. Galor (2012) surveys this literature and describes five possible mechanisms: (1) the rise in the level of parental income, which increased the opportunity cost of raising children and promoted investment in "quality" rather than "quantity" (Becker 1960; Becker and Lewis 1974); (2) the rise in the demand for human capital, which promoted a similar change in investment from quantity to quality (Galor and Weil 1999; Galor and Moav 2002); (3) the decline in infant and child mortality; (4) the decline in the gender gap (Galor and Weil 1996); and (5) the decline in the relative importance of children as "old-age security" with the development of new saving opportunities in the capital markets. According to this study, the arrival of railroads increased the socioeconomic occupation score and literacy rates, results which are consistent with mechanisms (1) and (2). I also find a positive effect of railroads on the age of marriage, which might imply a different quantity-quality tradeoff: adults invest in their own human capital, delay marriage and because of that have fewer children.

The second relevant strand of the literature is the historical debate regarding the relationship between industrialization and human capital. While some historians and economists have argued that human capital was not an important factor during the Industrial Revolution (Landes, 2003), more recent studies found complementarities between industrialization and different aspects of human capital, especially for later periods (Feldman and Van der Beek, 2016; Pleijt, Nuvolari and Weisdorf, 2016; Franck and Galor, 2017). Katz and Margo (2013) argue that the manufacturing labor force in the United States "hollowed out" during the second half of the 19th century, as the demand for middle-skilled artisans declined while that for low- and high-skilled jobs increased. In line with the later studies, I find that the economic development induced by railroads had

a large positive effect on human capital and on the occupational structure, implying that the net effect of the hollowing-out process described by Katz and Margo (2013) was positive in the case of railroads.

The third relevant strand of the literature include studies that estimate the effects of human capital on fertility (Becker, Cinnirella and Woessmann, 2009; Murphy, 2010; Klemp and Weisdorf, 2010; Bleakly and Lange, 2009) or the effect of economic development on fertility (Franck and Galor, 2015; Wanamaker, 2012). This literature is trying to empirically examine different mechanisms that might explain the Demographic Transition. Wanamaker (2012) studies textile mills in South Carolina between 1880 and 1900 and finds a substantial negative effect of the mills on fertility, similar to the effect I find.

Galor and Mountford (2008) combine the discussion of the Demographic Transition, the increase in trade and the Great Divergence. They argue that the effect of economic development induced by trade on fertility and human capital might have been different in different regions, because trade increased specialization. In regions that specialized in skilled-intensive industrial goods the gains from trade were translated to more human capital and fertility decreased, while in regions that specialized in unskilled-intensive agricultural goods the gains from trade were translated to increased fertility. In line with this theory, I find that both effects on fertility and human capital were significantly larger in counties that were initially more developed, and significantly lower in less developed counties. I also show that counties that were initially more developed increases their specialization in skilled-intensive industries due to the arrival of railroads.

The fourth strand of the literature which is relevant for this paper include studies of the effect of railroads and other transportation infrastructures, many of them focus on 19th century United States. There is a long-running debate in the literature over the role of railroads in the economic growth of the United States during this period. Taylor (1951) argued that the railroads advanced economic

growth, while Fishlow (1965) claimed that the railroad played a more passive role and its growth was driven by economic development. The recent literature tends to support Taylor's side. Using an identification strategy similar to the one presented in this paper, Atack, Haines and Margo (2008) show that railroads contributed to the rise of large factories and the decline of small artisans during the second half of the 19th century. Atack, Bateman, Haines and Margo (2010) find that railroads had no effect on population density, but did affect the trend of urbanization during that period. Donaldson and Hornbeck (2016) use a different approach, based on Trade Theory, to show that railroads increased market access and had a large effect on the value of agricultural land and on general welfare. Papers that discuss the effects of railroads in other countries, such as Banerjee, Duflo and Qian (2012), Hornung (2015), Berger and Enflo (2017) and Donaldson (2018), usually find substantial effects on trade, incomes, urbanization, population density, industrialization and the level of GDP per capita.

One important difference between the methodology used in this paper and the rest of the railroad literature is the use of a dynamic instrument for railroads, which allows for county fixed effects. In most other studies that use straight connecting lines as an instrument for railroads the lines are fixed in time, and those studies also don't control for the distance to the nodes of the network, which their location might be endogenous. Another difference is the outcome variables analyzed. Other railroad studies either focus on short-term effects, or show a long-term effect on GDP or urbanization without providing much evidence for the mechanism behind it. This study creates a link between the railroad literature and the long-term growth literature, by providing evidence for the effect of railroads on the basic ingredients of long-term growth: The Demographic Transition and the accumulation of human capital. Third, this study also shows heterogeneity in the effect of railroads in different regions, as mentioned above. While the effect of railroads on economic development found in this study is in line with the modern railroad literature, the effect I find on

manufacturing is smaller relative to what other studies find, and I also don't find an effect on urbanization. I discuss those differences in detail in section 5.1.

3. Descriptive analysis: Railroads, Economic Development, Fertility and Human Capital in 19th Century US

3.1 Scope and Data

Most of the data is taken from the decennial censuses, the Agricultural Census and the Manufacturing Census carried out by the US Bureau of the Census Library throughout the 19th century. As in the case of other historical databases, the data is far from perfect. For example, in the "Remarks on the Tables of Manufacturing Industry" in the 1870 survey, the author describes differences in the methodologies used in the manufacturing surveys of 1860 and 1870, such as the exclusion of the mining industry in 1870 which is partly compensated for by the inclusion of the milling of ores. Another example is the unavailability of the population in certain age groups in some of the years, which creates inconsistency in the measures of survival fertility. While little can be done to correct these deficiencies, it is worth noting that the main results of the paper are based on a panel analysis which includes fixed effects for counties and years. These fixed effects are likely to capture most of the inconsistencies between different years or between the different methods used by the assistant marshals responsible for collecting the data in each county.

The county-level data was published by the National Historical Geographic Information System (NHGIS), which also publishes geocoded county boundaries for each period.[2] The data for population and the location of cities was published by the U.S. Census Bureau and Erik Steiner, as a part of the Spatial History Project of the Center for Spatial and Textual Analysis at Stanford University.[3] I also use individual-level data published by IPUMS –

[2] https://www.nhgis.org/
[3] https://github.com/cestastanford/historical-us-city-populations

USA, including full-count data for 1850, 1880 and 1910, which allows me to compute some of the main variables.[4] Railroad data was published by the "Railroads and the Making of Modern America" project of the Center for Digital Research in the Humanities at University of Nebraska–Lincoln.[5] The CPI measure used to calculate real variables is based on the work of Lawrence H. Officer and Samuel H. Williamson, in "The Annual Consumer Price Index for the United States, 1774-2014".[6]

The sample period is 1850-1910. Data limitations regarding some of the variables prevented me from going back further than 1850. Reasons for stopping in 1910 include WWI, which is considered a "structural break" between the 19th century and the 20th century by many historians (see for example Hobsbawm, 2010), and the wide spread of automobiles after 1910, which transformed transportation in the US and probably affected the instrument used in this paper.

The analysis is carried out at the county level, and most of the results are limited to 1,490 counties east of the 95° line of longitude whose boundaries remained unchanged during the period. I used only those counties because most of the western counties were sparsely populated at the time (see Figure 2), the boundaries of the western counties changed during 1850-1910, and the empirical strategy makes less sense for counties far away from the largest cities. However, the list of major cities used to construct the connecting lines includes San Francisco, because railroads directed to San Francisco crossed many of the counties in the sample. According to a sensitivity analysis the results are robust for using other boundaries instead of the 95° line of longitude, or when western counties whose boundaries remained unchanged are also included in the sample.

[4] https://usa.ipums.org/usa/
[5] http://railroads.unl.edu/
[6] https://www.measuringworth.com/uscpi/

Figure2 : Population Density (individuals per km^2), 1880

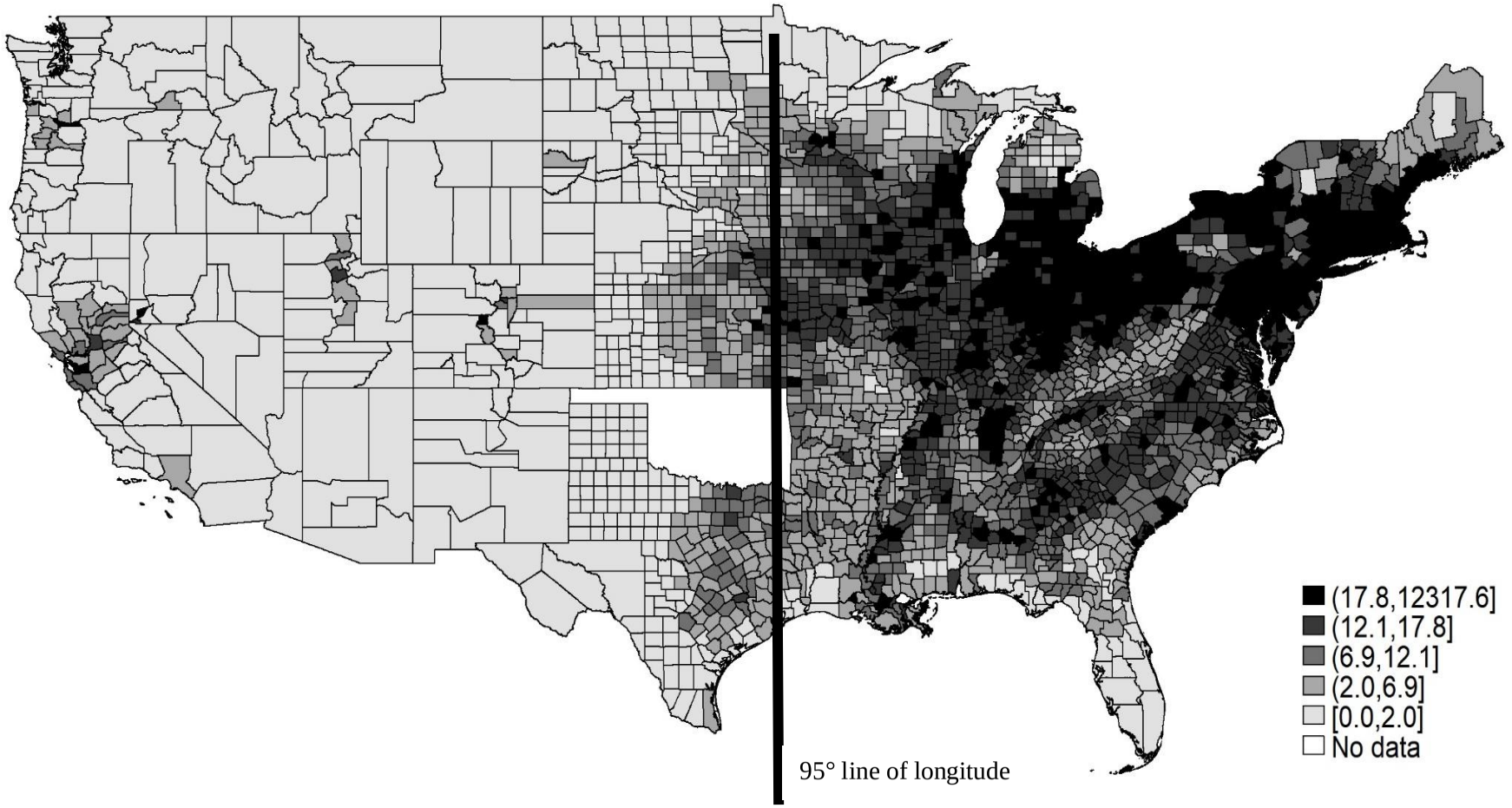


Table 1 : Variables Definitions

Variable	Definition	1850*	1860	1870	1880*	1890	1900	1910*
Distance to Railways	Airline distance in kilometers between the centroid of each county and the nearest railway	V	V	V	V	V	V	V
Duncan's (1961) Socioeconomic Index	Mean occupational socioeconomic score for males aged 25-64. The score is based on income and education in each occupation in 1950	V			V			V
Share of Non-Agriculture Workers	Males aged 16-65 employed in non-agriculture industries / males aged 16-65	V			V			V
Real Value of Manufacturing Output Per Capita	Real value of manufacturing output / population, 1850 prices	V	V	V	V	V	V	
Survival Fertility	Children aged ~5-19 / females aged ~20-44	V	V	V	V	V	V	V
Total Fertility Rate	Total fertility rate of females aged 15-44	V			V			V
Adult Males Literacy	1 - % illitirate adult males aged 20+	V		V	V		V	V

Notes: The stars (*) represent years in which I use IPUMS individual-level full count data. In other periods I used data which is originally aggregated at the county level. For survival fertility in some of the years the ages of children or adults are a bit different due to data limitation, and in some years I use males instead of females for the same reason.

Table 2: Summary Statistics

	Variable	Mean	p25	Median	p75	Standard Deviation	Change in Mean Since 1850
1850	Distance to Railways (km)	64.08	10.40	35.83	85.58	80.12	
	Duncan's (1961) Socioeconomic Index	19.43	17.33	18.69	20.52	3.54	
	Share of Non-Agriculture Workers	41%	26%	38%	54%	20%	
	Real Value of Manufacturing Output Per Capita	24.67	3.11	10.01	30.10	40.23	
	Survival Fertility	2.88	2.67	2.94	3.16	0.42	
	Total Fertility Rate	5.30	4.44	5.30	6.19	1.44	
	Adult Males Literacy	89%	84%	92%	96%	10%	
1910	Distance to Railways (km)	4.45	0.85	2.28	5.25	6.24	-93%
	Duncan's (1961) Socioeconomic Index	20.53	18.00	19.99	22.27	3.70	6%
	Share of Non-Agriculture Workers	50%	35%	46%	63%	21%	20%
	Real Value of Manufacturing Output Per Capita (1900)	92.94	19.85	44.85	114.90	122.20	277%
	Survival Fertility	1.76	1.45	1.77	2.09	0.38	-39%
	Total Fertility Rate	4.42	3.51	4.36	5.27	1.12	-17%
	Adult Males Literacy	87%	80%	89%	96%	11%	-3%

Notes: The data is based on 1,490 counties east of the meridian 95° west longitude line. The averages are at the county level and do not represent the average for all of the United States. The real value of manufacturing output per capita is calculated for 1900 instead of 1910 due to lack of data, and it is calculated according to 1850 prices.

The following sections describe the main variables used in the research. Table 1 presents the definitions of the variables and the years for which they are available. The variables that are calculated using the full count data published by IPUMS are available only for 1850, 1880 and 1910, and the literacy and manufacturing variables are also available only for some of the years. Table 2 presents summary statistics for an average county in 1850 and 1910, and the change in means between the years.

3.2 Transportation Infrastructure

Transportation infrastructures in the United States during the early years of the nation were relatively limited. The first river steamboats and canals started to operate in the beginning of the 19th century, and the construction of the Erie Canal in 1817 spawned a boom of canal-building around the country. Over 3,326 miles of man-made waterways were constructed between 1816 and 1840 (Cowan 1997). Towns located along major canal routes became major industrial and trade centers, while exuberant canal-building pushed some states to the brink of bankruptcy. The National Road (also known as the Cumberland Road), built between 1811 and 1837, was another important early transportation infrastructure, connecting the Potomac and Ohio Rivers and serving as a main transport path to the West. However, after the middle of the 19th century the focus started to shift from canals and roads to the newest and most exciting technology: railroads.

The first railroad steam locomotive in the United States, the "Stourbridge Lion", was imported from the UK in 1829, and operated in Honesdale, Pennsylvania. A domestic locomotive manufacturing industry was established during the 1830's and grew rapidly since then. The first common carrier railroad in the United States, The Baltimore and Ohio Railroad, opened in 1830, and others soon followed. In 1840, the railroad mileage in the United States was already similar to that of canals, by 1850 it exceeded that of canals by more than two to one, and by 1860 the United States had more miles of railroad than the rest of

the world combined (Atack, Bateman, Haines and Margo, 2010). The First Transcontinental Railroad that reached San Francisco Bay was opened in 1969, and by the beginning of the 20th century a dense network of railroads covered most of the United States. Figure 3 presents the railroads network in 1850, 1880 and 1910. In 1850 most of the railroads were located in the Northeast and they only started to expand westwards, while in 1910 most of the country was covered by a dense network of railroads. The new transportation infrastructures were usually built in undeveloped areas.

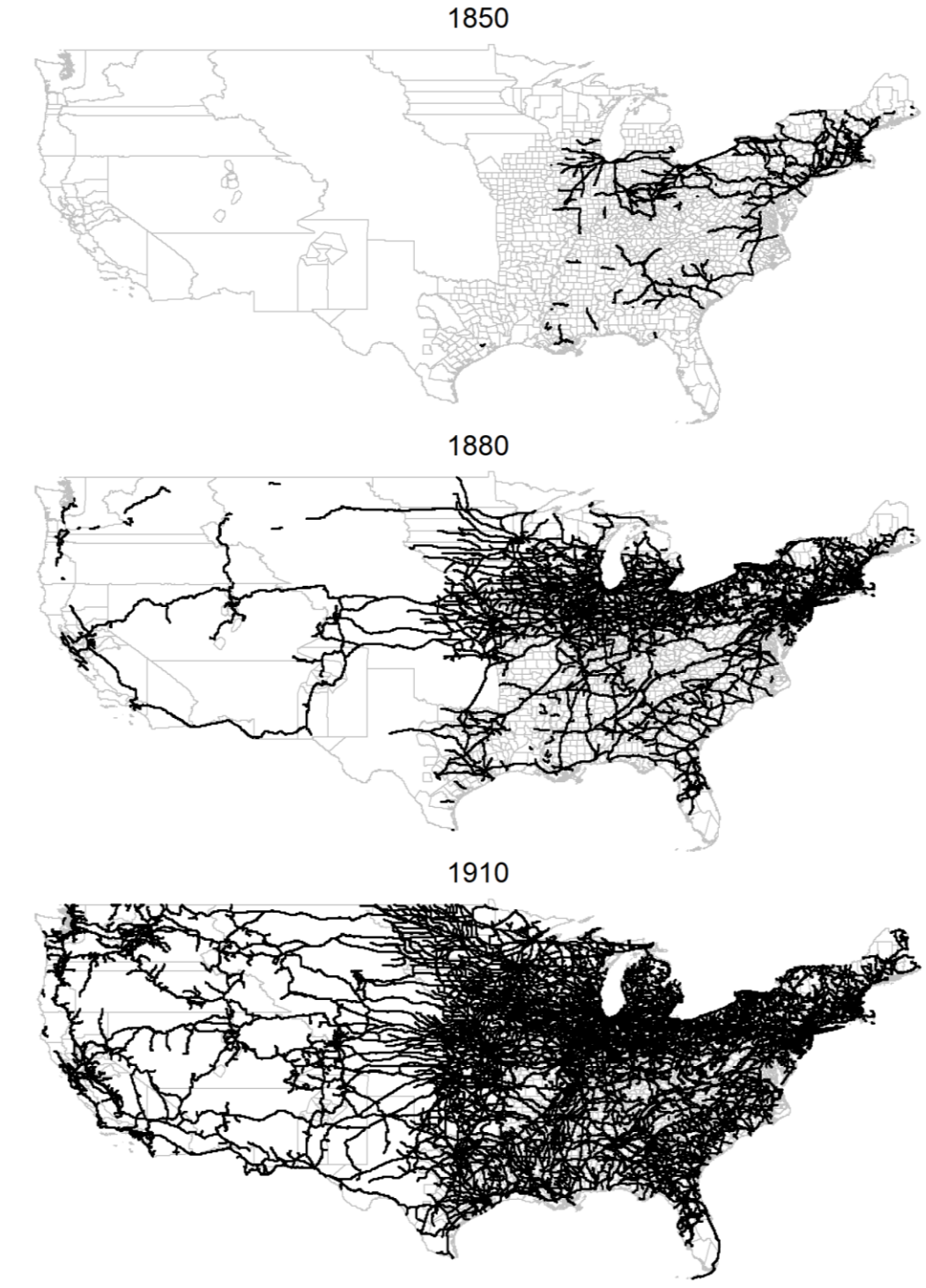
As can be seen in Table 2, for the sample of counties used in this study, the average distance between the centroid of a county and the nearest railroad was about 64 kilometers in 1850, compared to less than 5 kilometers in 1910. Those numbers represent a major improvement in transportation costs. For example, according to the 1932 Atlas of the Historical Geography of the United States, in 1800 it took more than 6 weeks to get from New York to the future location of Chicago, and by 1830 the new canals shortened the journey to about 3 weeks. By 1857 railroads shortened it to only two days, and by 1930 trains made this distance in less than a day.[7]

3.3 Economic Development

This study considers three different aspects of economic development relevant to the 19th century. The first one is the development of the manufacturing sector, measured by the real value of manufacturing output per capita. The US manufacturing sector was established in the Northeast in the end of the 18th century. During the period 1838-1880 the number of steam engines used for manufacturing in the United States increased from 1,420 to 56,123, while the number of waterwheels and turbines increased in a much more moderated pace, from 29,324 to 55,404 (Rosenberg and Trajtenberg, 2004). The relative share of the US in the world manufacturing output grew from 0.8% in 1800 to 7.2%

[7] Nice maps from the atlas are available here: http://dsl.richmond.edu/historicalatlas/

Figure 3: Railroads in 1850, 1880 and 1910



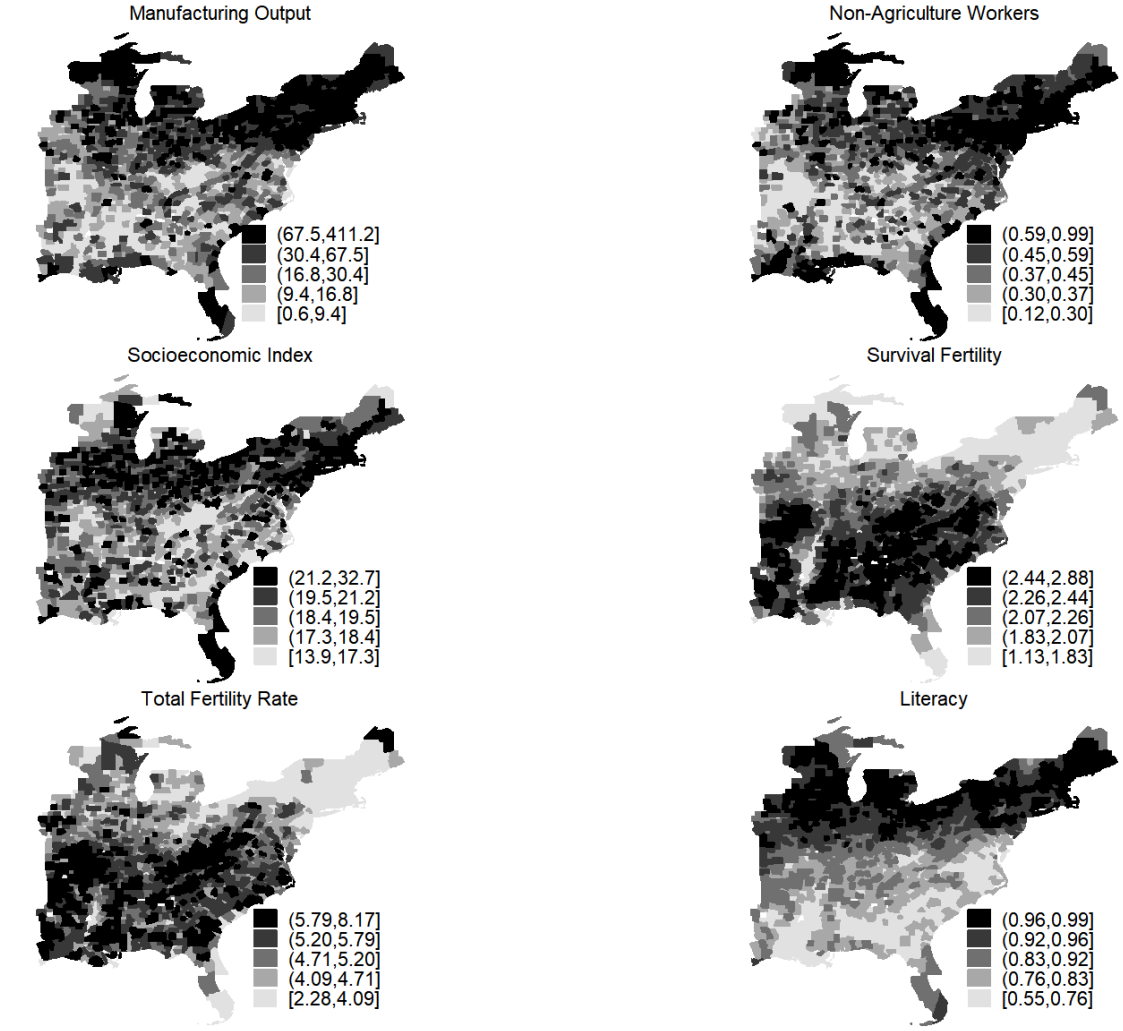
in 1860, then to 14.7% by 1880, and by 1900 the US passed the UK and became the largest manufacturing power in the world, producing 23.6% of the world manufacturing output (Kennedy, 2010). This rapid increase can be seen in Table 2: the average real value of manufacturing output per capita in a county almost tripled between 1850 and 1900.

Other industrialization variables, such as the value of capital invested in manufacturing per capita and the share of males employed in manufacturing, produce very similar geographic distribution and time trends as the real value of manufacturing output per capita. The correlations between those three different measures of the manufacturing sector are about 0.9. The value of manufacturing output was chosen as the main measure of industrialization in this study, because reports from the 19th century cast doubts on the consistency of the manufacturing capital definitions and data, and because the share of males employed in manufacturing is similar to another variable we use, the share of non-agricultural male workers. For 1870 there is also data on water wheels and steam engines, which are used in other papers as a measure of industrialization (Franck and Galor 2017; Pleijt, Nuvolari and Weisdorf 2016). The correlation between the value of manufacturing output per capita in this year and the horse power of water wheels is 0.92, and for the horse power of steam engines it is 0.95.

Figure 4 presents the geographic distribution of the manufacturing output per capita and the other main variables, averaged over all the periods. The most industrialized part of the US in 1850 was the Northeast. During the period it expanded towards the Midwest, and later also to more southern counties. There is a large geographic variation in the average level of industrialization: the average value of manufacturing output per capita in the Northeast is about 6 times larger than the average value in the South.

The second aspect of economic development analyzed in this study is the industry structure of the labor market, captured by the share of non-agricultural

Figure 4: Descriptive Maps, Averages for All Periods



Notes: The maps show the average value for each of the main outcomes for all the periods for which each variable exists. See Table 1 for the periods available for each variable.

workers. As can be seen in Table 2, in an average county the share of adult males not employed in agriculture increased from 41% in 1850 to 50% in 1910. I focus on non-agricultural workers instead of manufacturing workers because the share of manufacturing workers was relatively small in many of the counties, and because the rise of the services sector was an important driver of the increasing demand to human capital in this period (and is somewhat ignored by

the literature). According to Figure 4, there is a large geographic variation in the distribution of industries: in an average county in the Northeast 67% of the males were not employed in agriculture, while in the South only 38%.

However, the movement out of agriculture is only part of the story. Economic development led to higher demand for many occupations that are characterized by high levels of human capital and income, including teachers, engineers, lawyers, doctors etc. The third aspect of economic development analyzed in this paper is the occupation structure of the economy, as captured by Duncan's Socioeconomic Index for occupations (Duncan, 1961). The index is based on the education and income of individuals in different occupations, according to a survey held in 1947. Using this index, I assume that the ranking of occupations did not change significantly between 1850 and 1947. This could be a reasonable assumption for some occupations, such as Lawyers, Physicians and unskilled laborers, but probably not for all of them. Studies indicated that measures of occupational standing could be problematic for the research of inter-generational occupational mobility or gender differences, especially if the measures are based on much later data.[8] Because of that I also use the share of individuals above or below some cutoff, and not only the index itself. There are several other indexes for occupations available in the data, but all of them are based on income or education of workers with those occupations in 1950, due to lack of data from earlier years.[9] The correlations between the different measures are between 0.8 and 0.9 and using them in the analysis produces very similar results to using Duncan's Socioeconomic Index.

Table 3 presents rankings and several other characteristics for the 40 most common occupations in 1880. The top occupations include lawyers, physicians,

[8] See a discussion and some relevant papers here: https://usa.ipums.org/usa/chapter4/sei_note.shtml
[9] A description of the different measures and a discussion regarding the differences between them can be found here: https://usa.ipums.org/usa/chapter4/chapter4.shtml#OCCSTANDING

Table 3: Characteristics of the Top 40 Most Frequent Occupations, 1880

Occupation	Duncan's Socioeconomic Index	Share in population	Mean Age	Female Share	Literate Share	Frequency
Lawyers and judges	93	0.4%	39.3	0.2%	99.9%	67,593
Physicians and surgeons	92	0.5%	42.9	2.4%	99.9%	84,906
Teachers	72	1.3%	27.4	68.6%	99.9%	230,507
Managers, officials, and proprietors	68	3.9%	40.8	5.3%	99.7%	699,428
Compositors and typesetters	52	0.3%	28.3	4.0%	100.0%	61,088
Clergymen	52	0.4%	44.8	0.3%	99.7%	64,445
Bookkeepers	51	0.3%	31.5	4.9%	100.0%	61,718
Salesmen and sales clerks	47	2.3%	27.1	8.1%	99.9%	418,837
Stationary engineers	47	0.3%	36.4	0.2%	99.6%	58,121
Milliners	46	0.2%	29.5	97.3%	100.0%	40,807
Clerical and kindred workers	44	0.6%	29.9	4.9%	99.9%	108,159
Tinsmiths, coppersmiths, and sheet metal workers	33	0.2%	33.3	0.2%	99.8%	38,861
Machinists	33	0.5%	34.9	0.2%	99.9%	86,230
Craftsmen and kindred workers	32	0.6%	37.9	0.3%	99.3%	99,964
Meat cutters, except slaughter and packing house	29	0.4%	34.1	0.3%	99.6%	74,428
Brickmasons, stonemasons, and tile setters	27	0.6%	41.0	0.1%	99.0%	102,792
Dressmakers and seamstresses, except factory	23	1.3%	28.6	99.5%	99.5%	242,274
Tailors and tailoresses	23	0.7%	36.4	37.7%	99.5%	121,074
Bakers	22	0.2%	34.1	2.8%	99.6%	37,626
Millers, grain, flour, feed, etc.	19	0.2%	40.7	0.3%	99.4%	44,223
Housekeepers, private household	19	0.4%	34.3	99.2%	98.2%	73,158
Carpenters	19	2.2%	40.9	0.1%	99.3%	393,178
Operative and kindred workers	18	7.7%	29.6	22.8%	98.9%	1,395,059
Barbers, beauticians, and manicurists	17	0.3%	30.6	6.4%	98.9%	45,570
Sailors and deck hands	16	0.4%	33.5	0.7%	98.2%	69,582
Painters, construction and maintenance	16	0.7%	34.2	0.3%	99.7%	117,321
Blacksmiths	16	1.0%	38.4	0.1%	98.8%	172,392
Cooks, except private household	15	0.6%	32.7	73.6%	91.4%	106,971
Truck and tractor drivers	15	0.9%	33.9	0.2%	98.1%	162,042
Farmers (owners and tenants)	14	25.2%	40.9	1.8%	97.5%	4,543,949
Molders, metal	12	0.2%	32.4	0.1%	99.7%	39,714
Laundressses, private household	12	0.6%	36.5	98.6%	89.4%	108,743
Gardeners, except farm, and groundskeepers	11	0.2%	44.9	1.6%	97.8%	38,653
Mine operatives and laborers	10	1.5%	32.9	0.2%	98.1%	274,464
Fishermen and oystermen	10	0.2%	34.7	0.4%	96.8%	41,773
Hucksters and peddlers	8	0.3%	37.7	5.5%	98.5%	55,564
Laborers	8	11.5%	33.1	7.2%	94.3%	2,079,835
Private household workers	7	5.4%	24.7	86.1%	96.1%	970,872
Farm laborers, wage workers	6	18.1%	23.1	14.7%	93.5%	3,252,112
Lumbermen, raftsmen, and woodchoppers	4	0.2%	d	0.4%	96.9%	42,790

Notes: The table presents the 40 most frequent occupations in 1880. Duncan's socioeconomic index is based on education and income level for each occupation in the middle of the 20th century (Duncan, 1961). The other variables are calculated using the full-count IPUMS data base for 1880.

teachers and managers, while lumbermen and farm laborers are at the bottom. The correlation between the index, which is based on 1950's data, and the share of literate adults in each occupation, which is based on contemporary data, is

0.43. According to Figure 4, there is a geographic variation in the score, but it is small relative to the variation in other variables: in an average county in the Northeast the score is 21.59, while in the South it is 18.53. The same is true for the time variation in the index, which is smaller than the changes in other variables, as can be seen in Table 2.

3.4 Fertility and Human Capital

Fertility is measured in this study in two ways: the number of children aged 5-18 per women aged 20-44 (i.e. survival fertility, as measured by Fernández, 2014), and the total fertility rate of women aged 15-44. Survival fertility is available for more period than the total fertility rate, and using only surviving children above age 5 eliminates most of the effect of changes in infant mortality on birth decisions taken by parents (Haines, 1998). The age definitions for survival fertility changes slightly for some of the years because of data limitation, but due to the inclusion of year fixed effects in the econometric model this is not a problem for the analysis. Total fertility rate is calculated using the full-count data files for 1850, 1880 and 1910, and can be affected by the trends in mortality rate during the period. As we shall see, the results are similar for both measures.

According to Table 2, the number of children per adult declined by 39% in the average county between 1850 and 1910, while total fertility rate declined by 17%. Figure 4 presents the geographical distribution of both measures. As can be seen, the regional differences in fertility were large. In an average county in the Northeast there were 1.7 children for each women, compared to more than 2.3 children per women in an average county in the South. Looking at Figure 4, one can also see the strong negative correlation between both measures of fertility and our measures of economic development.

While the occupational index capture some aspects of human capital, the main measure we use for it in this study is adult male literacy rates. The United States

was a highly literate society: in 1840 more than 90% of white adults in the US were literate, a level similar to those in Scotland and Germany and higher than those in England and France (Fishlow, 1966). According to Table 2, literacy rates in an average county declined between 1850 and 1910. This result is also true at the country-level, and it appears in other data sources and other studies (see for example Hazan, 2009). It could be a result of the mass immigration to the US during the period. Figure 4 presents the geographic distribution of literacy during 1850-1910: in an average Northeastern county about 95% of adult males were literate, compared to about 78% of the adult males in an average Southern county. Looking at Figure 4, one can see the strong positive correlation between literacy and our measures of economic development.

One shortcoming of using literacy rate as an outcome variable, is that in many counties it was close to 100% already in 1850. Thus, the effect reported on literacy is probably smaller than the real effect of economic development on human capital. A robustness analysis presented in the following sections excludes counties close to 100% literacy rates, and, as expected, provides larger estimates for the effect of railroads on literacy.

3.5. The Relationship Between Economic Development, Fertility and Literacy

While this paper empirically considers economic development, fertility and literacy as outcomes of railroads, in the theoretical chain of reactions economic development is a mediator for the effect, while literacy and fertility are the "final outcomes". A railroad can directly affect the economic development variables by lowering transfer costs, but its effect on fertility and literacy is probably not direct and works through the effect on economic development.

In this section we will focus on the second part of this chain, and analyze the relationship between economic development, fertility and literacy without considering the railroads. The connecting lines which are used as instruments

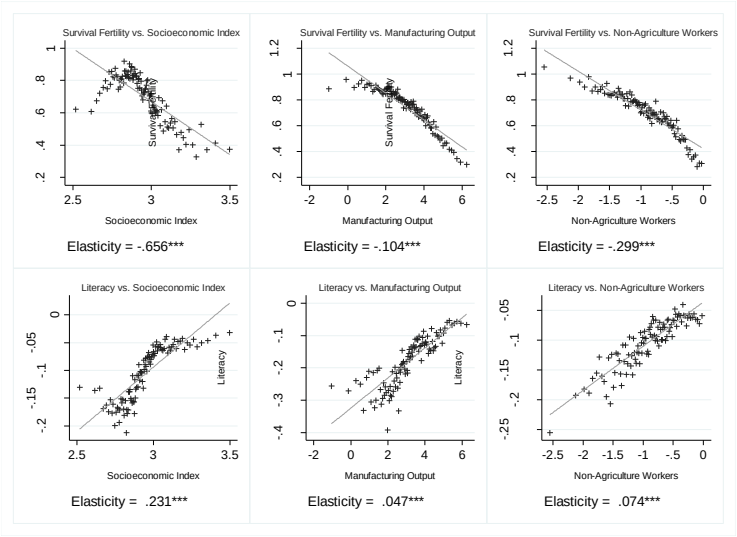
for railroads cannot be used as instruments for the economic development variables, since we don't know what the channel is, and the exclusion restriction does not hold. Because of that, we cannot identify the direction of the causality. However, we can at least control for some of the unobservables using county and year fixed effects.

Figure 5 presents the effect of economic development on survival fertility and literacy. Panel A presents the unconditional effect, while Panel B presents the effect after controlling for county and year fixed effects. All the variables are logged. The figure also reports the elasticities between the variables. It seems that the log-linear trend line fits the data better after controlling for fixed effects. All the effects are highly significant but one: the effect of manufacturing output value on literacy in Panel B. As we shall see in the following sections, the manufacturing sector in the US included some industries that did not require high levels of human capital, a fact that might explain this result.

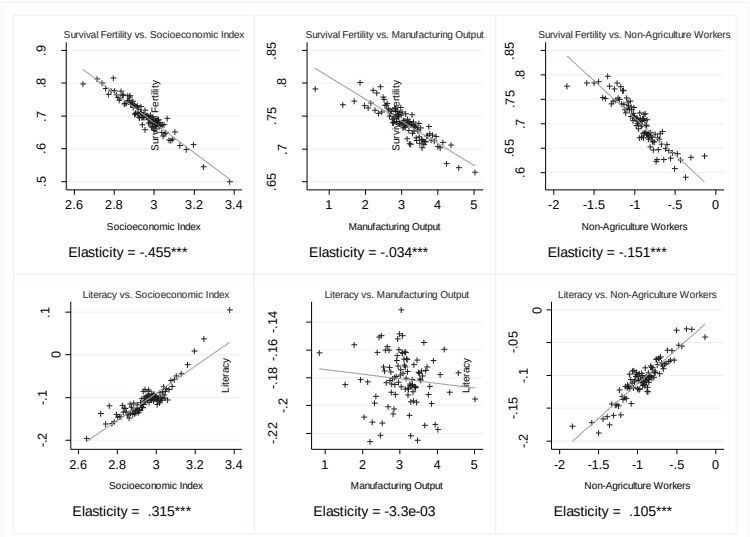
Table 4 presents a calculation of the size of the conditional effect. For example, if a county starts in the 25th percentile in respect to the socioeconomic index, with a value of 15.96, and increases the index to the 75th percentile value of 20.04, and if survival fertility was at the median level of 1.72, it will decrease by 0.2. For each woman there will be 0.2 less children. This moves a county from the median level of fertility to the bottom 25^{th}. The effects of manufacturing and the share of non-agriculture workers on fertility are even larger, and the effects of the socioeconomic index and the share of non-agriculture workers on literacy are also large. Looking at the trends in economic development and fertility between 1850 and 1910, those coefficients imply that the increase in socioeconomic index can account for 12% of the decrease in fertility, the increase in manufacturing output can account for 43% of the decrease in fertility and the increase in the share of non-agriculture workers can account for 14% of the decrease in fertility. Of course, those large effects might also reflect reverse causality or omitted variables.

Figure 5: The Relationship Between Economic Development, Survival Fertility and Literacy

Panel A: Unconditional Relationship



Panel B: Relationship Conditional on County and Year Fixed Effects



Notes: The county-year observations are grouped into 100 equal-sized bins, each represented by a "+" sign. All variables are logged. Standard errors are clustered at the county level. The stars represent the significance of the elasticities: *** p<0.01, ** p<0.05, * p<0.1.

Table 4: The Size of the Effect of Economic Development on Survival Fertility and Literacy

Effects conditional on county and year fixed effects

Outcome variable	Estimated coefficients		Variables distribution			Movement from p25 to p75	
	Survival Fertility	Literacy	p25 in 1880	p50 in 1880	p75 in 1880	Absolute change to median fertility	Absolute change to median literacy
Socioeconomic Index	-0.455	0.315	15.96	17.94	20.04	-0.20	8%
Share of Non-Agriculture Workers	-0.151	0.105	27%	37%	53%	-0.26	10%
Manufacturing Output Per Capita	-0.034	0	7.46	18.90	50.10	-0.33	0%
Survival Fertility			1.93	1.72	1.49		
Adult Males Literacy			94%	98%	99%		

3.6. The Relationship Between Distance to Railroads and the Outcomes

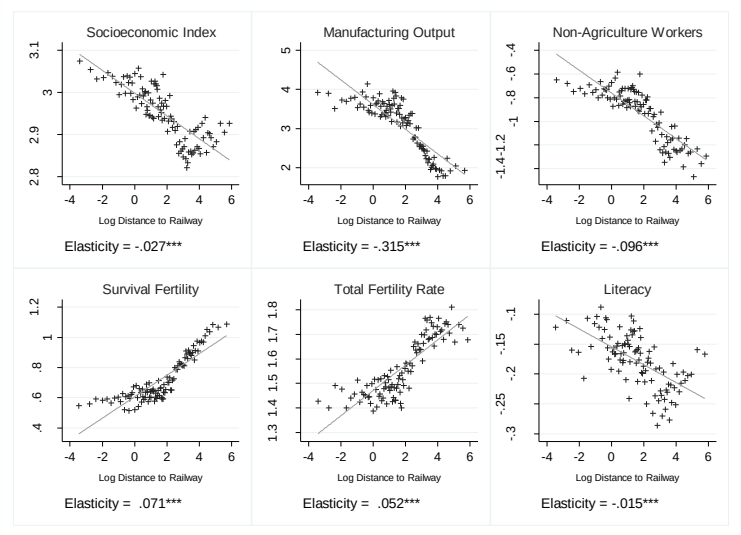
This section provides a basic analysis of the effect of railroad on the main variables of interest, without using any instrument. While this analysis may be biased because of the endogeneity of the location and timing of railroads construction, it is still interesting to see the general patterns in the data, without focusing on the counties that are affected by our natural experiment and drive the IV results.

Figure 6 presents the correlation between the distance to railroads and the outcomes. Panel A presents the unconditional relationship between the variables, while Panel B presents the effect after controlling for county and year fixed effects. All the variables are logged. According to both panels there is a clear positive correlation between the distance and fertility measures, and clear negative correlations between the distance and our measures for economic development and human capital. The figure also reports the elasticities between the variables, which are all highly significant. According to panel B, decreasing the distance to the nearest railroad by 10% will increase the socioeconomic index by 0.14%, increase the value of manufacturing output per capita by 0.7%, increase the share of non-agriculture workers by 0.55%, reduce surviving fertility by 0.068%, reduce the total fertility rate by 0.081%, and increase literacy by 0.097%.

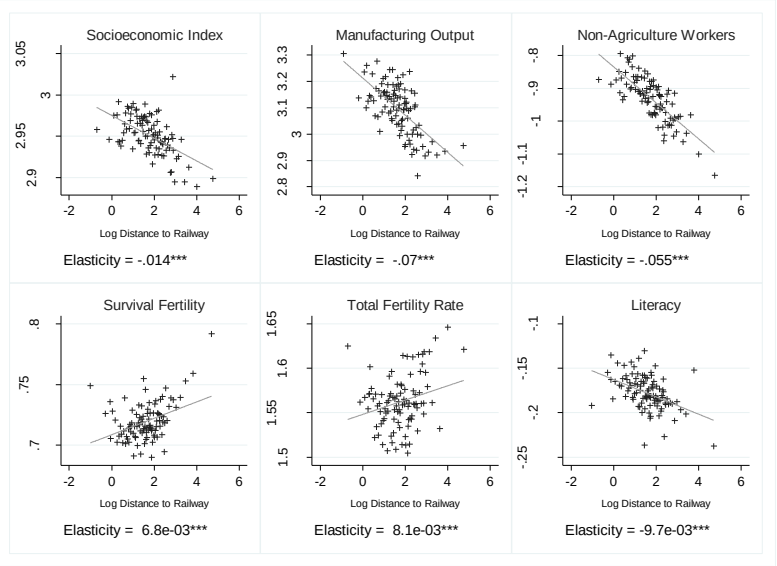
Figure 7 analyzes the trends in outcomes before and after the arrival of railroads, which is defined as the year in which the distance between the centroid of a county and the nearest railroad was smaller than 10 km. Panel A shows an unconditional version of the analysis, while Panel B presents the residuals for the outcomes after controlling for fixed effects for counties and years. The figure also reports the coefficients for the time trends before and after the arrival of railroads. Even without using any specific natural experiment, it seems that economic development came after the railroads, and did not precede the railroads. According to Panel A only in survival fertility and literacy we see similar trends before and after treatment. The trend in the occupational socioeconomic index is negative prior to the arrival of railroads, the share of non-agricultural workers and the total fertility rate shows no trend before, and the trend in the value of manufacturing output is positive but much smaller than the trend after the arrival of railroads. The results are even stronger once we control for county fixed effects and year fixed effects: before the arrival of railroads there are no trends for the share of non-agricultural workers, the value of manufacturing output and both fertility measures, and there is a negative trend in the occupational socioeconomic index. Only in literacy we see a trend prior to the arrival of railroads, which is larger than the trend after the arrival of railroads, but this might reflect the fact that literacy rates are bounded by 100%. The results of this basic analysis strengthen the view that the arrival of railroads was an exogenous event in many counties, even without using any specific natural experiment. This implies that the elasticities presented in Figure 6 might represent a causal effect of railroads and are not biased due to reverse causality. The empirical strategy presented in the following sections will focus on a specific exogenous variation in the distance to railroads, based on the growth of new major cities, and as we shall see the results imply an even larger effect of railroads in this case.

Figure 6: The Correlation Between Distance to Railroads and the Outcomes

Panel A: unconditional



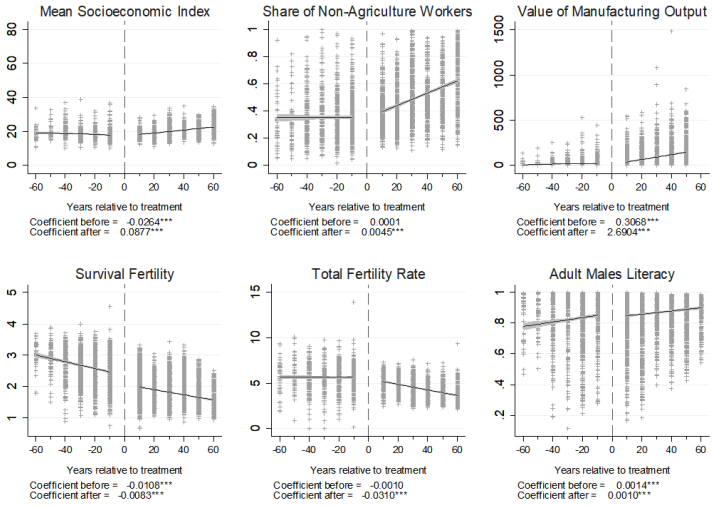
Panel B: Conditional on Fixed Effects for Counties and Years



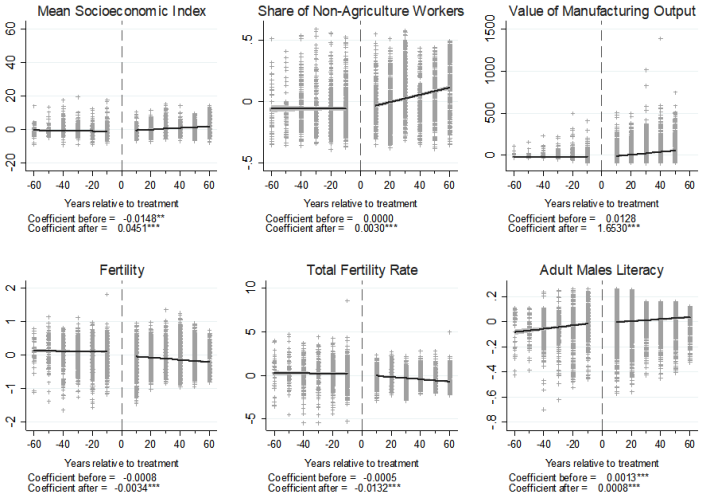
Notes: The county-year observations are grouped into 100 equal-sized bins, each represented by a "+" sign. All variables are logged. Standard errors are clustered at the county level. The stars represent the significance of the elasticities: *** p<0.01, ** p<0.05, * p<0.1.

Figure 7: Trends Before and After the Arrival of Railroads

Panel A: unconditional



Panel B: Conditional on Fixed Effects for Counties and Years



Notes: Each dot in the graphs is a county-year observation. Treatment in both panels is defined for each county as the year when the distance to railroad was below 10 km. According to this definition, 25% of the counties were already treated in 1850, while 90% of the counties were treated until 1910. In Panel B the outcomes are the residuals after controlling for fixed effects for counties and years. The figures also include a 95% confidence interval, but it is hard to see it due to the scale. The coefficients presented below each figure are for the trend lines before and after the treatment. The stars represent significance: *** p<0.01, ** p<0.05, * p<0.1.

4. Empirical Strategy

4.1 The General Framework

The identification strategy is best illustrated using the example presented in Figure 1. St. Louis, Cincinnati and Chicago experienced rapid growth during the second half of the 19th century, which led to the development of the transportation infrastructures that connected them to each other and to other major cities. The exact routes of the transportation infrastructures might be endogenous, but due to cost considerations their routs resembled straight connecting lines between the cities. Thus, a network of straight connecting lines might capture the exogenous part of the railroad network. The exclusion restriction assumption in this case is that after controlling the distance to the nearest major city, county fixed effects and year fixed effects, the distance to the connecting lines affects economic development only through its effect on the possibility that a railroad was built along this line.

The distance between US counties and connecting lines between large cities changed during the second half of the 19th century, thanks to booming new cities such as Chicago, Buffalo, Cleveland and Detroit, which functioned as transportation hubs. New railroads were built to transport goods between the new cities and older ones in the east, and the "middle counties" in between benefitted from the transportation infrastructure. Table 4 presents the top 10 most populated cities in 1850 and 1910. The population of all cities increased dramatically during the period, but the new industrial cities grew much faster than the older cities.

The empirical strategy involves two main choices: choosing the major cities and choosing how to draw the straight lines that connects them.

Using small cities for our purpose is problematic, since many of them appeared because of the railroads, and the traffic volume between them was small and probably did not affect the middle counties. One natural selection mechanism

Table 4: Top 10 Most Populated US Cities, 1850 and 1910

	1850		1910	
Rank	City	Residents	City	Residents
1	New York City	515,547	New York City	4,800,000
2	Baltimore	169,054	Chicago	2,200,000
3	Boston	136,881	Philadelphia	1,500,000
4	Philadelphia	121,376	St. Louis	687,029
5	New Orleans	116,375	Boston	670,585
6	Cincinnati	115,435	Cleveland	560,663
7	St. Louis	77,860	Baltimore	558,485
8	Albany	50,763	Pittsburgh	533,905
9	Pittsburgh	46,601	Detroit	465,766
10	Louisville	43,194	Buffalo	423,715

Notes: The table does not include cities which became neighborhoods in other cities. The network of straight connecting line used to construct the instrument also include cities that don't appear in this table, such as Sun Francisco, since they entered the top-10 list after 1850 and left the list before 1910.

for major cities would be to select all cities above some threshold of population size. However, this mechanism is problematic due to the dramatic increase in urbanization during the period. For example, a threshold of 40,000 residents produces 14 cities for 1850, 44 cities for 1880, 93 cities for 1900 and 138 cities for 1910, which most of them can be hardly considered as "major". An alternative mechanism is to start with the top X most populated cities, and in each period add to the list all the new cities that make it to the top X. For example, if X=10 the initial list of cities appears in Table 4 in the column of 1850; in 1860 Chicago, Buffalo and Newark enters the list; in 1870 San Francisco enters the list (almost the same year when the transcontinental railroad opened); in 1880 Cleveland enters the list and in 1910 Detroit enters the list. Cities can only enter the top X list, they do not leave the list, because it is not likely that railroads leading to a city will disappear just because it's rank decreased from the 9^{th} place to the 12^{th} place. Using this mechanism, the number

of major cities doesn't change dramatically during the years, and we can be confident that we are considering major cities. In most of the following analysis X will be equal to 10, but robustness tests include some alternatives. One small modification of this mechanism involves cities that became neighborhoods of other cities during the period 1850-1910: besides disappearing from the data set, these cities were also very close to other major cities, so there is no point in drawing a line to connect them. Therefore, the following cities were omitted for all time periods: Brooklyn, Spring Garden, Northern Liberties and Kensington.

The second choice we need to make using this identification strategy is how to construct the network of lines that connects the major cities. Using actual railroads data for that, as done in Hornung (2015), is problematic because the timing of railroad construction might be endogenous. I propose two different algorithms for constructing the network. The first and simplest one is to draw all possible lines between all the major cities in each period. This means that our network will also include somewhat "unrealistic" connecting lines, for example between New York and San Francisco, or between Buffalo and New Orleans. For each new major city that enters the list we add lines to all other cities. We will call this algorithm "all-lines" in short.

The second algorithm starts with a Minimum Spanning Tree (MST) for all the major cities in 1850, constructed according to Kruskal's Algorithm (Kruskal 1956). The algorithm identifies the minimum number of edges that connect all major cities, subject to the minimization of the total network distance. After 1850, for each new major city that enters the list we don't run the algorithm again, because this will eliminate some of the previous lines, which is unrealistic in respect to railroads. Instead, we just add one line between the new major city and the nearest major city that was included in the list in the previous period. We will call this algorithm "MST" in short. While most of the analysis is done on the sample of counties east of the 95 line of longitude, both algorithms

consider all the cities in the US, including western cities, because railroads leading to those cities passed through our sample counties.

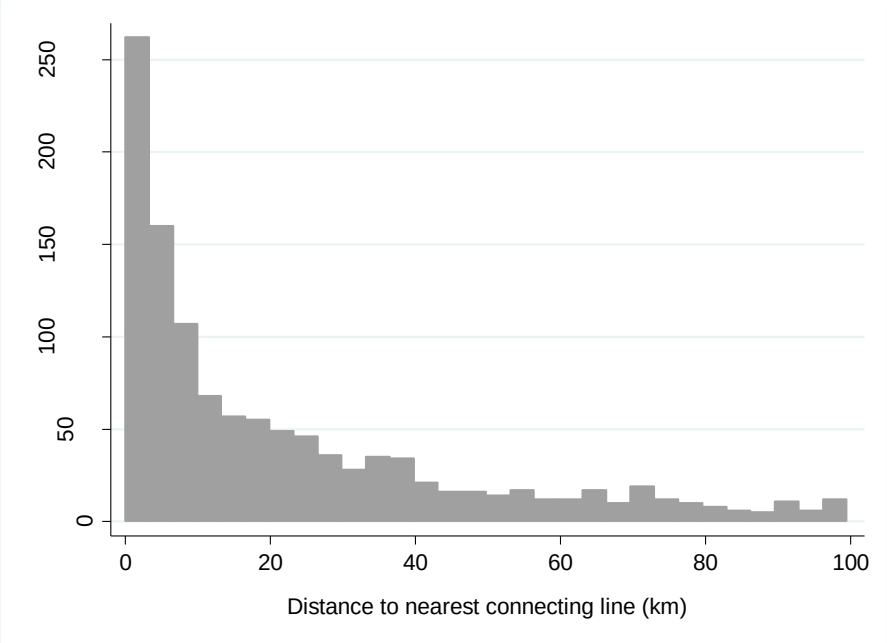
Both approaches have advantages and disadvantages. Comparing to the actual railroad network, the all-lines algorithm produces too many lines, while the MST algorithm produces too few lines. Both algorithms focus on the major cities and do not produce lines to remote counties, and for both of them the distance to lines is highly correlated with the distance to railroads. Figure A1 in online appendix A presents maps of lines and cities for both algorithms, for the case of 10 major cities, for the years 1850, 1880 and 1910. As the MST version is more sensitive to different specifications since it includes less lines. Therefore, most of the analysis is based on the all-lines algorithm, and the MST version is used to establish robustness.

Figure 8 presents a histogram of the distances to connecting lines in 1880, for using 10 major cities and the all-lines algorithm, for counties with distance shorter than 100 km (78% of the counties in the sample). As can be seen, there is a lot of variation, and many counties are in the range of 10-80 kilometers from the nearest connecting line.

Other methods for constructing the connecting lines are also possible. For example, we could take into account geographic barriers, the composition of industries in each city, existing canals and navigable rivers and so on, and construct a changing network of straight lines that resembles the railroad network and might produce a stronger first stage. We could also use different weights for different connecting lines, for example according to the size of the cities in both ends of the line, or we can use the distances to many connecting lines instead of only to the nearest one. However, the main advantage of the algorithms used here is their relative simplicity. The construction of the instrument does not involve making any complicated decisions along the way, and the algorithms is based on only one parameter: the number of major cities. Because of that we are not overfitting the real development of the railroads, and

Figure 8: Histogram of the distance to connecting lines, 1880

All-Lines Algorithm, Top 10 Major Cities, Counties With Distances < 100 km



it is more reasonable to argue that our instrument is as good as randomly assigned.

Several other papers use similar identification strategies to study the effect of transportation networks. Atack, Haines and Margo (2008), who also study railroads in 19th century US, use straight lines drawn between urban areas in 1820 and the closest major coastal port as an instrument for the existence of railroads crossing counties in 1850, and they also use information on the starting and endpoints of railroad engineering surveys authorized by Congress as an instrument for the existence of railroads crossing counties in the Midwest. Banerjee Duflo and Qian (2012) use the distance to the nearest straight line connecting historical cities in China as an instrument to the location of railroads. An important difference between the strategies used in those studies and the one used in this paper is that the list of major cities changes between 1850 and 1910, so the instrument presented here is dynamic. More straight lines are added for

each new city that enters the list. The dynamic nature of this natural experiment allows controlling for unobservables using county fixed effects (as well as year fixed effects). To mitigate concerns regarding the endogenous location of the major cities, I also control the distance to the nearest major city. To the best knowledge of the author, the only study which uses a dynamic instrument based on straight lines to estimate the effect of railroads is Hornung (2015), who study the effect of railroads in 19th century Prussia on the development of cities. Hornung (2015) included fixed effects for the cities he studies, but he adds new straight lines each time a new railroad is constructed. A mentioned before, this could be problematic since the timing of construction might be endogenous. In this study straight lines are added once a city enters the top-10 list, and I am not using any railroad data to construct the instrument.

4.2 First-Stage and Reduced Form

In this section, I show that the distance between the centroid of a county and the nearest straight line connecting two large cities has a strong first-stage relationship with the distance between the centroid of a county and the nearest railroad. Table 5 presents the results of the first-stage regressions. The full econometric model (column 3) is as follows:

(1) $Log\left(RAILDIST_{\mathrm{i},t}\right) = \beta_1 Log\left(LINEDIST_{\mathrm{i},t}\right) + \beta_2 Log\left(CITYDIST_{i,t}\right) + \delta_{\mathrm{i}} + \gamma_{\mathrm{t}} + \epsilon_{\mathrm{i,t}}$,

where $RAILDIST_{\mathrm{i},t}$ is the distance in year t between the centroid of county i and the nearest railroad, $LINEDIST_{\mathrm{i},t}$ is the distance between county i and the nearest connecting line between two of the 10 largest cities in any period $j \leq t$, $CITYDIST_{i,t}$ is the distance to the nearest major city in any period $j \leq t$, δ_{i} are county fixed effects and γ_{t} are year fixed effects. Throughout this paper, the standard errors are clustered at the county level, and all variables except dummies are in logarithm. Panel A of the table presents results for the all-lines algorithm, while panel B presets result for using the MST algorithm. Both

Table 5: First Stage – the Relationship Between the Distance to Connecting Lines and the Distance to Railroads

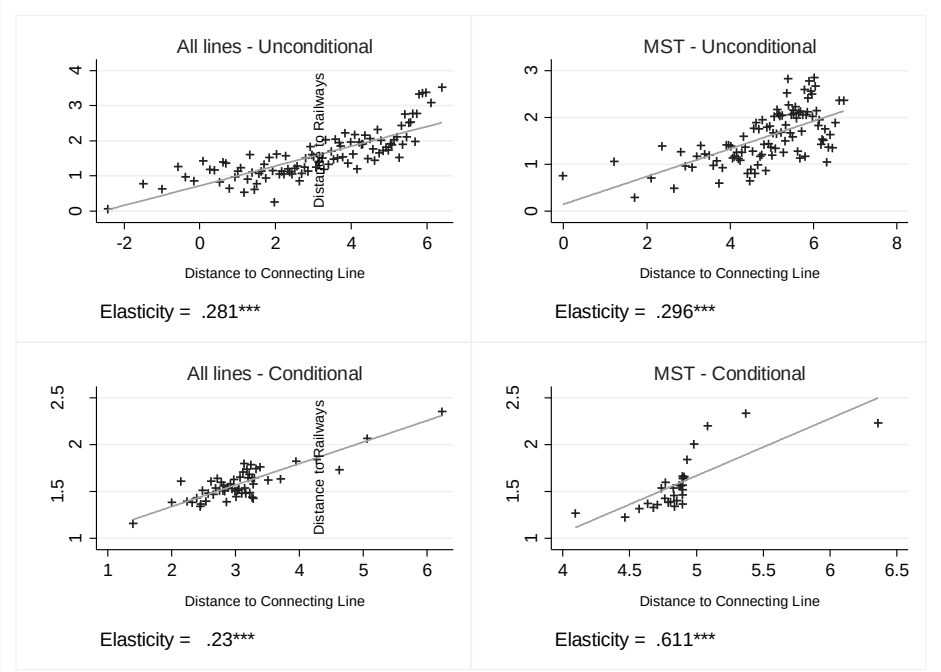
	(1) No Controls	(2) Including Fixed Effects	(3) Including Distance to Cities	(4) Including the West
Panel A: All-lines algorithm				
Distance to Lines	0.281***	0.203***	0.230***	0.258***
	(0.0174)	(0.0286)	(0.0292)	(0.0282)
Distance to Cities			-0.449***	0.0731
			(0.114)	(0.0887)
R-squared	0.087	0.499	0.502	0.519
F statistic	262.1	430.7	384	419
Panel B: MST algorithm				
Distance to Lines	0.296***	0.372***	0.611***	0.585***
	(0.0313)	(0.0748)	(0.0972)	(0.0980)
Distance to Cities			-0.798***	-0.278**
			(0.132)	(0.131)
R-squared	0.040	0.495	0.501	0.514
F statistic	89.60	427.7	384.1	419.1
Observations	10,395	10,395	10,395	11,039
Number of id		1,485	1,485	1,577
County Fixed Effects	no	yes	yes	yes
Year Fixed Effects	no	yes	yes	yes

Notes: All variables are in logarithm except the dummies. Standard errors are clustered at the county level. The stars represent significance: *** p<0.01, ** p<0.05, * p<0.1.

panels are based on the top-10 major cities in each period. Colum 1 shows the results without any controls, in column 2 we include only fixed effects, column 3 includes the distance to cities and column 4 provides results when including western counties for which the borders did not change much between 1850 and 1910. The main specification used in this paper is the one in column 3.

The results show a significant relationship between the distance to lines and distance to railroads for both versions of the instrument, also after controlling fixed effects and distance to cities. The coefficients can be interpreted as elasticities. For example, according to column 3 on panel A, reducing the distance to a connecting line by 10% will reduce the distance to the nearest railroad by 2.3%. The MST version provides larger elasticities, which are also

Figure 8: First Stage



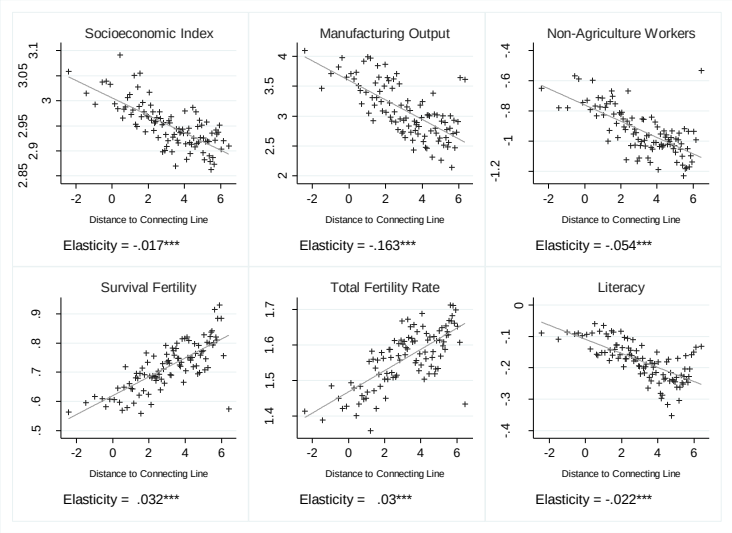
Notes: The county-year observations are grouped into 100 equal-sized bins, each represented by a "+" sign. All variables are logged. Standard errors are clustered at the county level. The stars represent the significance of the elasticities: *** p<0.01, ** p<0.05, * p<0.1.

more sensitive to the controls included and to other possible specifications, because of the fewer connecting lines.

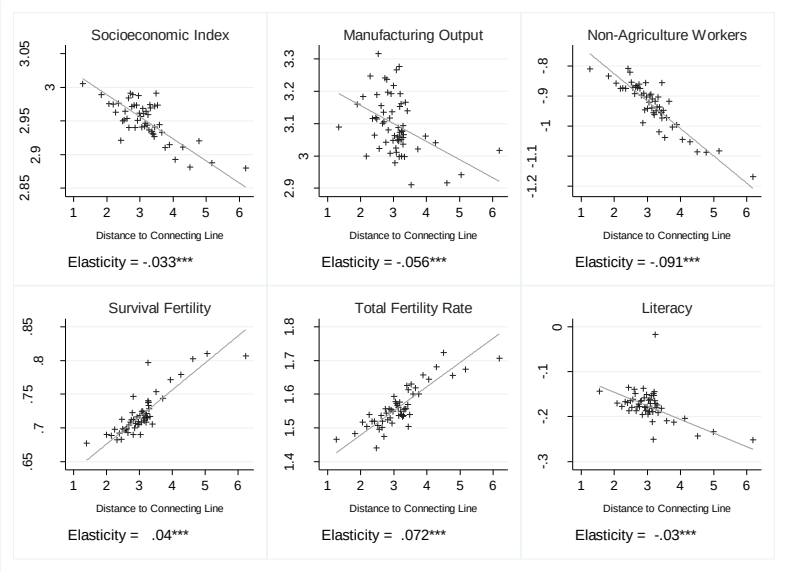
Figure 8 presents scatter plots for the first stage. The figure also reports the elasticities between the variables. The unconditional graphs clearly show significant correlations that are not driven by outliers (each dot represents about 90 county-year observations). However, the conditional graph of the MST version seems to be sensitive to some outliers, and it is less consistent with a log-linear model than the all-lines version. Scatter plots using the two algorithms with different number of major cities look very similar. Because of the sensitivity of the MST version, the rest of the results reported in this paper are based on the all-line version, and the MST version is used as a robustness exercise.

Figure 9: Reduced Form

Panel A: Unconditional Relationship



Panel B: Relationship Conditional on the Controls



Notes: The county-year observations are grouped into 100 equal-sized bins, each represented by a "+" sign. All variables are logged. Standard errors are clustered at the county level. The instrument is built using the all-lines algorithm and the top-10 major cities. Controls in panel B include the distance to the nearest major city and fixed effects for counties and years. The stars represent the significance of the elasticities: *** p<0.01, ** p<0.05, * p<0.1.

Figure 9 presents the reduced form, for the all-lines version of the instrument and top-10 major cities. The relationship between the instrument and the outcome variables is significant and is not driven by outliers. According to panel B, reducing the distance to the nearest connecting line by 10% increases the socioeconomic index by 0.3%, increases the value of manufacturing output by 0.56%, increases the share of non-agricultural workers by 0.9%, increases literacy by 0.3%, decreases survival fertility by 0.4% and decreases total fertility rate by 0.7%.

4.4 Pre-Treatment Differences

One concern is that counties along future connecting lines were already different from other counties prior to the appearance of new major cities and the railroads that connected them. If this is the case, the exclusion restriction assumption does not hold. As we mentioned in previous sections, historical evidence does not support this argument, since the new transportation infrastructures were usually built in undeveloped areas. Figure 7, discussed in previous sections, provide supportive empirical evidence for the exogeneity of railroads in general, without focusing on the natural experiment on which we base the instrument. Another way to address this concern, in the context of our natural experiment, is to create a binary version of the instrument, that separates the counties into two groups: a treatment group of counties that were far from the connecting lines in the "before period" and close to the connecting lines in the "after period", and a control group of counties that were far from the connecting lines in both the before period and the after period. More specifically, the analysis is done for the period 1850-1880, the treatment group include counties that their distance from the nearest connecting line was above the mean in 1850 and below the mean in 1880, and the control group include counties that were above the mean in both years. This definition provides 198 treatment counties and 642 control counties.

Figure 10 presents the results of the analysis. Panel A presents a map of the treatment and control counties, which also allows us to see some of the areas that drive the main results presented in the next sections. Panel B presents the outcomes for the treatment and control groups, before and after the treatment. In 1850 the treatment counties were actually less developed according to the socioeconomic occupation index, the share of non-agriculture workers and fertility measures, there was no difference in manufacturing output value between the treatment and control counties, and the literacy rates in the treatment counties were only slightly higher (the difference in literacy is not significant at 1% level). However, in 1880 we see a "reversal of fortunes": the treatment counties were significantly more developed in all aspects, except for the share of non-agriculture workers, where there is a large difference, but it is not precisely measured. According to these results, it seems highly unlikely that counties near future connecting lines were more developed before the appearance of the transportation infrastructures. These results are also robust to various definitions of the treatment and control groups.

Another way to address this concern is to regress the outcomes on both the current distances from connecting lines and the future ones, along with all the controls. Table 6 presents the results for the following econometric models:

(2) $Log(Y_{i,t}) = \beta_1 Log(LINEDIST_{i,t}) + \beta_2 Log(CITYDIST_{i,t}) + \delta_i + \gamma_t + \epsilon_{i,t}$,

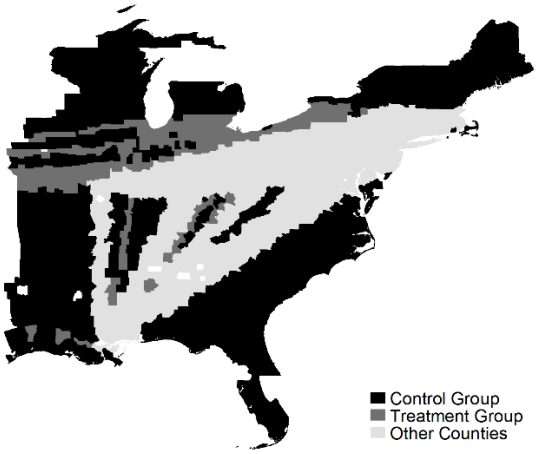
(3) $Log(Y_{i,t}) = \alpha_1 Log(LINEDIST_{i,t+20}) + \alpha_2 Log(CITYDIST_{i,t}) + \psi_i + \theta_t + u_{i,t}$,

(4) $Log(Y_{i,t}) = \lambda_1 Log(LINEDIST_{i,t}) + \lambda_2 Log(LINEDIST_{i,t+20}) + \lambda_3 Log(CITYDIST_{i,t}) + \varphi_i + \sigma_t + v_{i,t}$,

where $LINEDIST_{i,t+20}$ is the future minimum distance from the nearest connecting line in 20 years, and the other variables are as described above. The distances to currents connecting lines is highly correlated with the distance to future connecting lines, because many of the connecting lines were in place already in 1850, so the coefficients of $LINEDIST_{i,t+20}$ might be significant at

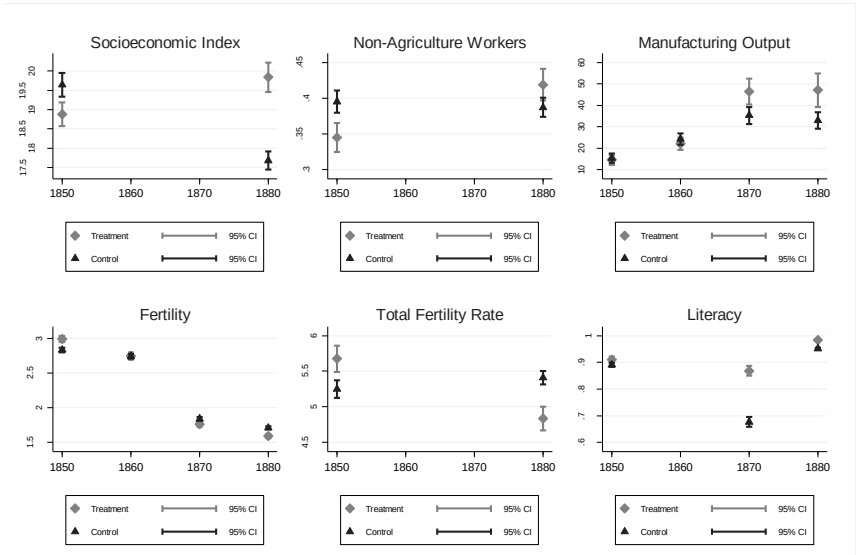
Figure 10: Pre-Treatment Differences

Panel A: A Map of the Treatment and Control Groups



Notes: The treatment group is defined as counties for which the distance to the lines in 1850 was larger than the mean distance, and the distance to the lines in 1880 was smaller than the mean distance. The control group is defined as counties for which the distance to the lines was larger than the mean distance in both 1850 and 1880.

Panel B: Outcomes Before and After Treatment



Notes: The treatment group is defined as counties for which the distance to the lines in 1850 was larger than the mean distance, and the distance to the lines in 1880 was smaller than the mean distance. The control group is defined as counties for which the distance to the lines was larger than the mean distance in both 1850 and 1880. The dots in the chart represent the mean of each group in each period. Data for the socioeconomic index, the share of non-agriculture workers and total fertility rate is not available for 1860 and 1870, data for literacy is not available in 1860.

Table 6: Current Connecting Lines vs. Future Connecting Lines

1850-1890

	(1) Mean Socioeconomic Index	(2) Share of Non-Agriculture Workers	(3) Value of Manufacturing Output	(4) Survival Fertility	(5) Total Fertility Rate	(6) Adult Males Literacy
Panel A: Separate Regressions						
Only Current Distances	-0.0334***	-0.0717***	-0.0677***	0.0353***	0.0529***	-0.0348***
	(0.00380)	(0.00927)	(0.0182)	(0.00275)	(0.00639)	(0.00338)
Only Future Distances	-0.0147	-0.00367	-0.101**	0.0270***	0.0304*	-0.0238
	(0.0135)	(0.0283)	(0.0447)	(0.00646)	(0.0181)	(0.0152)
Panel B: Combining Both Distances						
Current distances	-0.0348***	-0.0773***	-0.0650***	0.0347***	0.0546***	-0.0346***
	(0.00398)	(0.00970)	(0.0182)	(0.00273)	(0.00681)	(0.00357)
Future Distances	0.0176	0.0681**	-0.0820*	0.0153**	-0.0202	-0.00530
	(0.0126)	(0.0283)	(0.0441)	(0.00612)	(0.0153)	(0.0152)
Observations	2,964	2,965	7,057	7,419	2,960	4,452
Number of id	1,485	1,485	1,483	1,485	1,485	1,485
County Fixed Effects	yes	yes	yes	yes	yes	yes
Year Fixed Effects	yes	yes	yes	yes	yes	yes

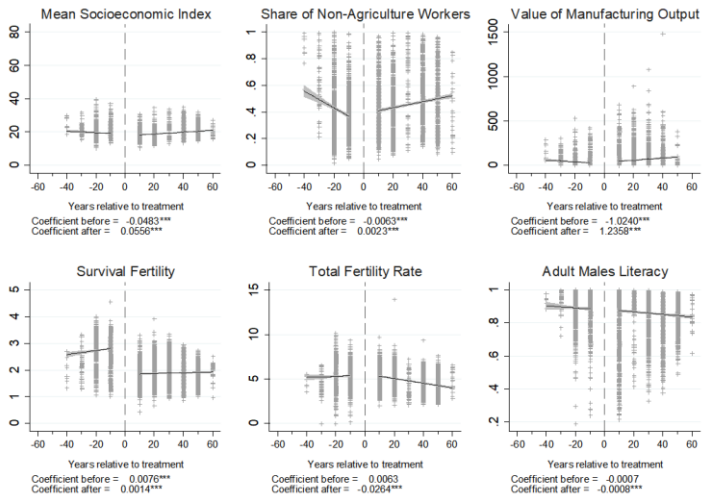
Notes: All variables are in logarithm except the dummies. Standard errors are clustered at the county level. Future distances are the distances in 20 years. The stars represent significance: *** p<0.01, ** p<0.05, * p<0.1.

least in some cases. But if the natural experiment is valid, the coefficients of $LINEDIST_{i,t}$ will reflect stronger correlations than the coefficients of $LINEDIST_{i,t+20}$. According to Table 6 this is indeed the case: the current distances are much more correlated with the current outcomes than the future distances, both when we run the regressions separately and when we combine the two distances in the same model.[10] Therefore, it seems likely that counties near the connecting lines, which are the ones driving the results, became more developed only after the growth of the connected cities.

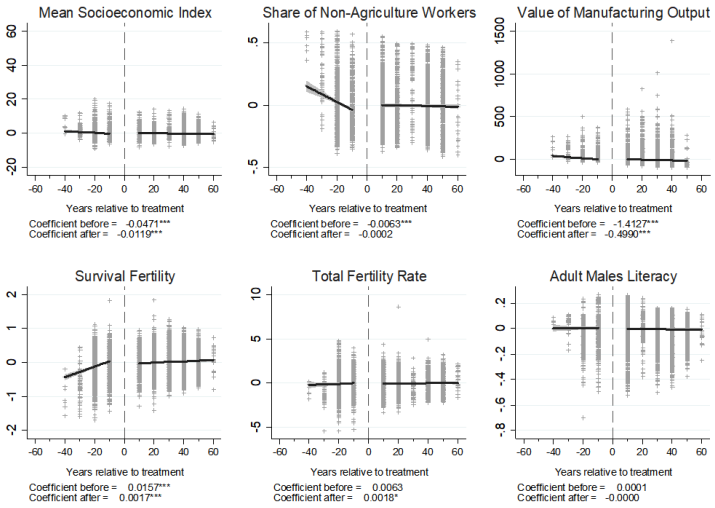
[10] This analysis can only be done for 1850-1890, since our sample is limited to 1850-1910, so the results presented in the table are a bit different than the results reported previously for the reduced form.

Figure 11: Trends Before and After the Arrival of Railroads, as Predicted by the First Stage

Panel A: unconditional



Panel B: Conditional on Fixed Effects for Counties and Years



Notes: Each dot in the graphs is a county-year observation. Treatment in both panels is defined for each county as the year when the predicted distance to railroad was below 10 km. In Panel B the outcomes are the residuals after controlling for fixed effects for counties and years. The figures also include a 95% confidence interval, but it is hard to see it due to the scale. The coefficients presented below each figure are for the trend lines before and after the treatment. The stars represent significance: *** p<0.01, ** p<0.05, * p<0.1.

Figure 11 presents the trends in outcomes before and after a "treatment", in a similar fashion to Figure 7. However, here the treatment is based on our natural experiment. The treatment year is defined as the year in which the predicted distance between the centroid of a county and the nearest railroad, as predicted by our first stage, was smaller than 10 km. Panel A shows an unconditional version of the analysis, while Panel B presents the residuals for the outcomes after controlling for fixed effects for counties and years. The figure also reports the coefficients for the time trends before and after the arrival of railroads. As can be seen in both the unconditional and the conditional version of this analysis, pre-treatment trends were not part of the story. Prior for getting close to a predicted railroad, the economic development and human capital variables either decline or don't show any trends, and fertility either increases or don't show any trends.

5. The Effect of Railroads on Economic Development, Fertility and Literacy

5.1 Main Results

Instrumenting the distance to railroads with the distance to connecting lines allows us to estimate the causal effect of railroads on economic development, fertility and human capital. Table 7 presents OLS and IV results for the main specification. The instrument used in this section is based on all the possible connecting lines between the 10 most populated cities in each period, while the MST version and other number of cities are presented in later sections as robustness tests. The econometric model is as follows:

$$(5)\quad Log(\mathrm{Y_{i,}}_{t}) = \beta_1 Log(RAILDIST_{\mathrm{i,}t}) + \beta_2 Log(CITYDIST_{i,t}) + \delta_{\mathrm{i}} + \gamma_{\mathrm{t}} + \epsilon_{\mathrm{i,t}}\,,$$

where $RAILDIST_{\mathrm{i,}t}$ is the distance to the nearest railroad, instrumented in panel A by the minimum distance to the nearest connecting line. The results establish a significant causal effect for the distance to railroads on different aspects of economic development, on both fertility measures and on adult male literacy.

Table 7: The Effect of Railroads on Economic Development, Fertility and Literacy

	(1) Mean Socioeconomic Index	(2) Share of Top 25% Occupations	(3) Share of Non-Agriculture Workers	(4) Value of Manufacturing Output	(5) Survival Fertility	(6) Total Fertility Rate	(7) Adult Male Literacy
Panel A: IV estimation							
Distance to Railways	-0.117***	-0.350***	-0.324***	-0.227***	0.175***	0.255***	-0.117***
	(0.0175)	(0.0491)	(0.0442)	(0.0726)	(0.0237)	(0.0369)	(0.0173)
Distance to Cities	-0.193***	-0.392***	-0.269***	-0.195***	0.102***	0.229***	-0.0516***
	(0.0183)	(0.0491)	(0.0442)	(0.0546)	(0.0225)	(0.0362)	(0.0150)
First stage F	65.68	65.62	65.79	69	61.61	65.88	60.52
Panel B: OLS estimation							
Distance to Railways	-0.0151***	-0.0713***	-0.0565***	-0.0708***	0.00708***	0.00941**	-0.00992***
	(0.00235)	(0.00640)	(0.00616)	(0.0113)	(0.00187)	(0.00462)	(0.00184)
Distance to Cities	3.916***	0.186	0.263*	3.653***	0.629***	0.692***	0.0690
	(0.0736)	(0.173)	(0.156)	(0.283)	(0.0646)	(0.119)	(0.0498)
Observations	4,449	4,446	4,450	8,540	10,389	4,445	7,422
Number of id	1,485	1,485	1,485	1,483	1,485	1,485	1,485
County Fixed Effects	yes	yes	yes	yes	yes	yes	yes
Year Fixed Effects	yes	yes	yes	yes	yes	yes	yes
Panel C: Size of the effect (according to the coefficients of Panel A)							
1. Geographic variation in 1880:							
Bottom 25% value in 1880	15.96	9%	27%	7.46	1.93	6.24	94%
Medial value in 1880	17.94	15%	37%	18.90	1.72	5.47	98%
Top 25% value in 1880	20.04	21%	53%	50.10	1.49	4.52	99%
Value if we start at the bottom 25% and decrease distance by 100%	17.83	13%	35%	9.15	1.59	4.70	105%
2. Time trend, 1850-1910:							
Change for an average county between 1850-1910	6%	-	20%	277%	-39%	-17%	-3%
Predicted change according to coefficient and change in the average distance to railway	11%	-	30%	21%	-16%	-23%	11%

Notes: In panels A and B all variables are in logarithm except the dummies. Standard errors are clustered at the county level. In panel C the variables are not logged. In the second part of panel C the analysis for the value of manufacturing output is done for 1850-1900, and not for 1850-1910, since no data is available for 1910. This analysis is not done on the share of top 25% occupations since this outcome is measured relative to the distribution in each time period. The stars represent significance: *** p<0.01, ** p<0.05, * p<0.1.

Reducing the distance by 10% increases the socioeconomic index by 1.17%, increases the share of non-agricultural workers by 3.24%, increases the value of manufacturing output by 2.27%, decreases survival fertility by 1.75%, decreases the total fertility rate by 2.55% and increases literacy by 1.17%. Because the socioeconomic index is based on 1950 data, an alternative measure is added to the outcome variables in column 2: the share of adult males in a county with an occupation index in the top 25% of the distribution of all the US adult males in that year. As can be seen in Table 3, the top occupations

according to the 1950 ranking were probably very similar to the top occupations in 1850. This variable increases by 3.5% due to a reduction of 10% in the distance to the nearest railroad.

Panel B of Table 7 presents OLS estimations. The OLS coefficients are significant and of the same sign as the IV coefficients, but they are biased towards zero relative to the IV coefficients. There could be several reasons for the difference between the IV estimates and the OLS estimates, including measurement errors and omitted variables, but the most probable reason is that the IV estimates are based on the main railroads between major cities, while the OLS estimates are based on all the railroads. It is not surprising to find that the distance to the main railroads has a larger effect on the outcomes than the distance to any railroad. However, our empirical strategy is not fit for measuring the effect of all the railroads, because we need to use major cities which didn't develop because of the railroads.

Panel C provides an analysis of the size of the effect of railroads on all the outcomes. This analysis includes two parts: comparing the effect to the distribution of the outcome variables in 1880, and comparing the effect to the trend in the outcome variables between 1850 and 1910. In the first part of the panel we can see some values for the geographic distribution of each outcome variable in 1880 (not logged), and the change in the value of the outcomes if a county starts at the bottom 25% and decrease the distance to railroad by 100%, calculated using the coefficients of panel A. For example if we start at a county with a socioeconomic score of 15.96, the bottom 25% line, and build a main railroad in the middle of that county, the score will increase to 17.83 – very close to the median, 17.94. The same exercise will increase share of non-agricultural workers almost to the median, reduce both fertility measures below the median, and increase literacy from 94% to above 100%. However, in respect to the value of manufacturing output the change is relatively small: an increase from 7.46 to 9.15, comparing to a median value of 18.9. Thus, we have a large

effect on fertility and literacy, a bit smaller effect on the socioeconomic index and the share of non-agricultural workers, and a small effect on the value of manufacturing output.

In the second part of panel C we can see the change in the mean value for each outcome between 1850 and 1910 (1900 in the case of the value of manufacturing output), and compare it to the predicted change according to the change in the mean distance between a county and the nearest railroad (93% decline, see Table 2). The predicted change for the socioeconomic index, the share of non-agricultural workers, total fertility rate and literacy is larger then the actual change, the predicted change in survival fertility is about 40% of the actual change, and the predicted change in the value of manufacturing output is only 8% of the actual change. This analysis imply that other events that happened between 1850 and 1910, such as the immigration waves or the Civil War, might have prevented the change that could have been induced by the railroads. However, this exercise is a bit problematic since our instrument is capturing the effect of main railroads and not the effect of any railroad.

How does these results compare to other papers? Wanamaker (2012), who studies South Carolina between 1880-1900, finds that each additional textile mill in a town reduces fertility by 6%-10%. We would expect that the effect of a new railroad will be larger than the effect of a single textile mill, and indeed, this is what we see according to the results above. For example, the predicted change between 1850 and 1910 in survival fertility due to the change in the mean distance to the nearest railroad is 16%.

Several papers in the railroads literature report estimates of the effect of railroads on variables similar to the ones examined here. Hornung (2015) finds no effect of railroads on fertility, measured as the ratio of children below age 5 to women aged 15-45, in Prussian cities between 1840-1871. The difference between his findings and the large effect reported here might be due to his focus only on urban dwellers. Many papers found an effect for railroads on different

aspects of economic growth, such is industrialization measures (Atack, Haines and Margo, 2008; Hornung, 2015; Berger and Enflo, 2017), urbanization (Atack Bateman Haines and Margo, 2010; Hornung, 2015; Berger and Enflo, 2017), income (Donaldson, 2018) or land prices (Donaldson and Hornbeck, 2016). While I find a medium effect on population density, reported in section 7, no effects on urbanization variables were found in this study. However, the available urbanization variables only include relatively large cities, and are equal to zero for most counties in most time periods. It is also important to note that the empirical strategy used in this paper is based mostly on the "middle counties", which were less urbanized, and include a control for the distance to the nearest major city. The large effects reported here on the distribution of occupations and industries probably represent at least some small-scale urbanization trends that are not captured by the available urbanization variables. Thus, the difference between this paper and other papers in the estimated effect on urbanization is probably a result of different empirical strategies and variables, and not an important contradiction.

A more important contradiction between this paper and the rest of the literature is the small and non-robust effect I find for railroads on the development of the manufacturing sector. This result also appears for industrialization measures used in other papers, such as the share of manufacturing workers. The manufacturing variables exists for most counties and time periods, and do not appear to be problematic like the urbanization variables. The difference in the results between this paper and other papers in this respect could be due to the use of a dynamic instrument and county fixed effects, which are absent in other papers that analyze the case of the US.

5.2 Robustness Tests

In this section we will discuss the sensitivity of the results to alternative specifications of the econometric model, to different samples and to alternative instruments.

Table 8 provides results for several different alternatives. Row a of the table presents the baseline results that appear in Table 7. Row b provide results for an econometric model that don't include any controls. Except for the value of manufacturing output, all of the coefficient in this case are smaller than in models that include fixed effects, like the baseline model and the model of row c. It seems that considering time trends and unobserved differences between counties increases the size of the effect on most outcomes. In row c we do not control for the distance to the nearest major city. The effect is larger than in the baseline results, implying that counties near major cities developed also because of the cities, and not just because of the railroads. Row d presents results for an econometric model that includes controls for the sex ratio, the share of whites and the share of foreign immigrants in each county. Those variables are not included in the main specification because they might also be outcomes of the railroads, so it is problematic to treat them as exogenous. Including the new controls in the regression decreases the effect of railroads on the economic development variables, but the effect on literacy doesn't change much and the effect on fertility increases a little. The effect on manufacturing, which was relatively small in the first place, becomes not significant in this case. Row e establishes that the results are not sensitive for controlling the distances to canals and navigable rivers, which were less important for economic development after 1850.

Row f of the table provide results for a different specification, in which we do not control for the distance to the nearest major city, but instead omit the major cities and the surrounding counties. The definition of the surrounding counties in this case is a radius of 70 kilometers, but other possibilities provide similar results. This specification provides larger coefficients than the baseline results, that are similar to the coefficients reported in row c. It could be that the cities influenced the development of counties that were relatively far from them. Rows g and h presents results for different cutoff longitude lines, instead of the 95-longitude line used in all the other results reported in the paper.

Table 8: Alternative Specifications for the Econometric Model and the Sample

Explenatory variable: distance to railways	(1) IV Mean Socioeconomic Index	(2) IV Share of Top 25% Occupations	(3) IV Share of Non-Agriculture Workers	(4) IV Value of Manufacturing Output	(5) IV Survival Fertility	(6) IV Total Fertility Rate	(7) IV Adult Males Literacy
a. Baseline specification	-0.117***	-0.350***	-0.324***	-0.227***	0.175***	0.255***	-0.117***
	(0.0175)	(0.0491)	(0.0442)	(0.0726)	(0.0237)	(0.0369)	(0.0173)
b. No controls	-0.0507***	-0.144***	-0.158***	-0.592***	0.115***	0.0878***	-0.0767***
	(0.00480)	(0.0151)	(0.0151)	(0.0535)	(0.00874)	(0.00772)	(0.00665)
c. Only fixed effects	-0.193***	-0.506***	-0.430***	-0.279***	0.205***	0.345***	-0.137***
	(0.0327)	(0.0818)	(0.0671)	(0.0817)	(0.0303)	(0.0569)	(0.0224)
d. Baseline specification + controls for sex ratio, share of whites and share of foreign born	-0.0819***	-0.241***	-0.227***	-0.117	0.176***	0.267***	-0.103***
	(0.0167)	(0.0439)	(0.0399)	(0.0773)	(0.0277)	(0.0437)	(0.0182)
e. Baseline specification + controls for distance to rivers and canals	-0.109***	-0.335***	-0.312***	-0.232***	0.178***	0.241***	-0.115***
	(0.0162)	(0.0463)	(0.0419)	(0.0735)	(0.0239)	(0.0340)	(0.0169)
f. No major cities and surrounding counties, not controling for distance to cities	-0.188***	-0.490***	-0.407***	-0.252***	0.205***	0.344***	-0.143***
	(0.0301)	(0.0759)	(0.0615)	(0.0793)	(0.0301)	(0.0543)	(0.0225)
g. 93 Longitude cutoff (instead of 95)	-0.179***	-0.514***	-0.443***	-0.409***	0.293***	0.399***	-0.174***
	(0.0401)	(0.109)	(0.0924)	(0.146)	(0.0693)	(0.0850)	(0.0385)
h. 97 Longitude cutoff (instead of 95)	-0.106***	-0.322***	-0.298***	-0.181***	0.157***	0.236***	-0.107***
	(0.0152)	(0.0425)	(0.0389)	(0.0647)	(0.0190)	(0.0317)	(0.0149)
i. Including the West	-0.119***	-0.352***	-0.309***	-0.177***	0.144***	0.253***	-0.107***
	(0.0186)	(0.0499)	(0.0439)	(0.0632)	(0.0173)	(0.0387)	(0.0151)
j. Including Civil War side X year interactions	-0.0116	-0.119***	-0.194***	-0.140*	0.0899***	0.129***	-0.0187**
	(0.00965)	(0.0307)	(0.0339)	(0.0714)	(0.0139)	(0.0236)	(0.00875)

Notes: All variables are in logarithm except the dummies. Standard errors are clustered at the county level. The stars represent significance: *** p<0.01, ** p<0.05, * p<0.1.

It seems that focusing on the more developed eastern counties increases the effect of railroads. In row i we include Western counties which their borders did not change much between 1850 and 1910. The results are very similar to the baseline results. Row j provide results for a model which allows for different time trends in the North and the South, defined according to the different sides of the Civil War. Such time trends might be a result of the different institutions and culture of the North and the South, or a direct result of the Civil War. Since most of the railroads were built in the North, including region-year interactions kills a large part of the overall variation in the change of the distance to the

nearest railroad, and the coefficients are smaller. However, most of the coefficients are still significant, and all of them have the same sign as in the baseline results. These results imply that counties which got access to railroads experienced fast economic development even when comparing to the trends in other counties in the same region.

The previous results are based on the all-lines version of the instrument and on the 10 largest cities in each period. As discussed in previous sections, there is no intrinsic reason for choosing the 10 largest cities, or for choosing one specific way of drawing the network of lines over the other. Thus, Table 9 presents a sensitivity analysis for different versions of the instrument. The versions include both different number of major cities, and using the MST version instead of the all-lines version of the network that connects all the cities.

Using less than 8 major cities produces a weak first stage, because several important cities in the Midwest fall out of the sample, and so are the connecting lines attached to them. On the other hand, using more than 25 cities also produces a weak first stage, since in this case in the earlier years the artificial network includes lines that did not yet existed. Looking at the coefficients of Panel A, it seems that most of the outcomes are robust to using different numbers of major cities. However, as in the previous analysis, the effect on the value of manufacturing output is less robust than the other outcomes, and in some cases it becomes non-significant.

The MST version in panel B provides similar coefficients to the all-lines version in general, but the first stage does not hold for 25 cities in this case. The all-lines version is more robust to different number of cities, since it includes more lines. This means that the decision on the number of major cities is less important in the case of the all-lines version of the instrument.

Table 9: Alternative Specifications for the Instrument

Explenatory variable: distance to railways	(1) IV Mean Socioeconomic Index	(2) IV Share of Top 25% Occupations	(3) IV Share of Non-Agriculture Workers	(4) IV Value of Manufacturing Output	(5) IV Survival Fertility	(6) IV Total Fertility Rate	(7) Adult Males Literacy
Panel A: All lines							
Top 8 cities	-0.115***	-0.350***	-0.325***	-0.0457	0.174***	0.234***	-0.0116
	(0.0172)	(0.0479)	(0.0438)	(0.0781)	(0.0269)	(0.0337)	(0.00994)
Top 9 cities	-0.117***	-0.345***	-0.324***	-0.211***	0.180***	0.260***	-0.110***
	(0.0181)	(0.0497)	(0.0448)	(0.0738)	(0.0247)	(0.0381)	(0.0162)
Top 10 cities	-0.117***	-0.350***	-0.324***	-0.227***	0.175***	0.255***	-0.117***
	(0.0175)	(0.0491)	(0.0442)	(0.0726)	(0.0237)	(0.0369)	(0.0173)
Top 11 cities	-0.121***	-0.361***	-0.341***	-0.196**	0.170***	0.258***	-0.116***
	(0.0189)	(0.0525)	(0.0482)	(0.0760)	(0.0235)	(0.0390)	(0.0182)
Top 12 cities	-0.0918***	-0.275***	-0.283***	-0.135*	0.153***	0.229***	-0.0888***
	(0.0151)	(0.0412)	(0.0400)	(0.0699)	(0.0207)	(0.0334)	(0.0144)
Top 15 cities	-0.0735***	-0.235***	-0.261***	-0.102	0.155***	0.214***	-0.0769***
	(0.0129)	(0.0359)	(0.0360)	(0.0716)	(0.0224)	(0.0299)	(0.0122)
Top 20 cities	-0.118***	-0.375***	-0.318***	-0.183	0.101*	0.357***	-0.0935**
	(0.0423)	(0.121)	(0.103)	(0.217)	(0.0535)	(0.105)	(0.0379)
Top 25 cities	-0.171***	-0.447***	-0.345***	-0.452	0.177*	0.444***	-0.0966**
	(0.0548)	(0.140)	(0.121)	(0.430)	(0.0990)	(0.132)	(0.0414)
Panel B: MST							
Top 8 cities	-0.139***	-0.355***	-0.297***	0.00509	0.241***	0.333***	-0.0427*
	(0.0355)	(0.0844)	(0.0743)	(0.126)	(0.0696)	(0.0897)	(0.0259)
Top 9 cities	-0.0893***	-0.271***	-0.228***	-0.167**	0.113***	0.180***	-0.0789***
	(0.0139)	(0.0398)	(0.0364)	(0.0665)	(0.0171)	(0.0303)	(0.0123)
Top 10 cities	-0.0951***	-0.287***	-0.245***	-0.183***	0.118***	0.198***	-0.0838***
	(0.0153)	(0.0434)	(0.0401)	(0.0688)	(0.0177)	(0.0340)	(0.0137)
Top 11 cities	-0.112***	-0.329***	-0.272***	-0.249***	0.132***	0.217***	-0.0919***
	(0.0165)	(0.0459)	(0.0407)	(0.0707)	(0.0181)	(0.0340)	(0.0139)
Top 12 cities	-0.100***	-0.304***	-0.271***	-0.255***	0.130***	0.215***	-0.0916***
	(0.0155)	(0.0449)	(0.0421)	(0.0734)	(0.0186)	(0.0352)	(0.0143)
Top 15 cities	-0.0855***	-0.267***	-0.246***	-0.184**	0.116***	0.201***	-0.0875***
	(0.0149)	(0.0439)	(0.0430)	(0.0762)	(0.0201)	(0.0366)	(0.0150)
Top 20 cities	-0.163***	-0.489***	-0.482***	-0.281*	0.225***	0.504***	-0.141***
	(0.0472)	(0.149)	(0.157)	(0.156)	(0.0600)	(0.142)	(0.0419)
Top 25 cities	-0.192	-0.260	-0.593	0.312	0.178*	0.657	-0.150*
	(0.162)	(0.346)	(0.490)	(0.336)	(0.0971)	(0.501)	(0.0896)

Notes: All variables are in logarithm except the dummies. Standard errors are clustered at the county level. The stars represent significance: *** p<0.01, ** p<0.05, * p<0.1.

Table 10: Effect in Less Literate Counties

Outcome variable = Literacy	(1) IV Literacy <= 1	(2) IV Literacy <= 0.98	(3) IV Literacy <= 0.96	(4) IV Literacy <= 0.94	(5) IV Literacy <= 0.92
Distance to Railways	-0.117***	-0.134***	-0.158***	-0.177***	-0.245***
	(0.0173)	(0.0205)	(0.0256)	(0.0352)	(0.0635)
Distance to Cities	-0.0516***	-0.0719***	-0.0620*	-0.0872**	-0.120
	(0.0150)	(0.0198)	(0.0330)	(0.0389)	(0.113)
Observations	7,422	6,386	5,321	4,333	3,671
R-squared	0.090	0.141	0.090	0.095	-0.284
Number of id	1,485	1,472	1,346	1,164	1,030
First stage F	60.52	57.79	47.77	30.10	15.30
County Fixed Effects	yes	yes	yes	yes	yes
Year Fixed Effects	yes	yes	yes	yes	yes

Notes: All variables are in logarithm except the dummies. Standard errors are clustered at the county level. The stars represent significance: *** p<0.01, ** p<0.05, * p<0.1.

Tables 8 and 9 establishes that the effect of railroads on most aspects of economic development, on fertility and on literacy is robust to different specifications of the econometric model, the sample group and the instrument. However, the effect of railroads on the value of manufacturing output is less robust than the effects on the other outcomes, and as we have seen before it is also relatively smaller. This is also true for other possible manufacturing variables, such as the share of males employed in manufacturing or the value of capital invested in manufacturing.

The last robustness test considers the effect on literacy rates. As mentioned before, literacy rates were close to 100% in many counties already in 1850. This implies that the effect we found for literacy is probably smaller than the real effect on human capital. Table 10 analyzes the effect on literacy in samples that include less literate counties. The first column presents the results we have seen before for all the counties, while in the rest of the columns the sample is restricted to less literate counties. As expected, the effect gets larger once we restrict the sample.

6. Heterogeneity in the Effect of Railroads

6.1. "The Small Divergence": Heterogeneity by Initial Development Level

In this section, I decompose the average effect described in the previous sections in order to determine whether it varies between different groups of counties. Specifically, I use the following econometric model:

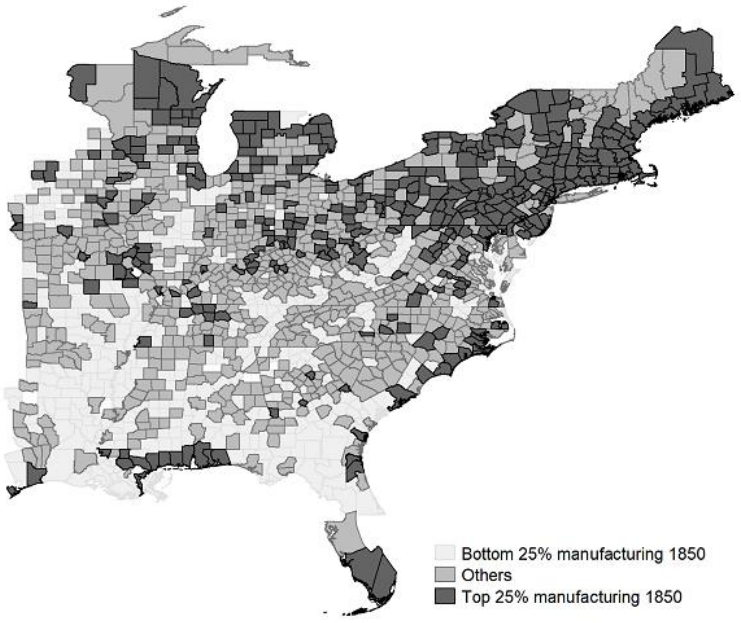
(6) $Log(\mathrm{Y}_{\mathrm{i},t}) = \beta_1 Log(RAILDIST_{\mathrm{i},t}) + \beta_2 GROUP_i * Log(RAILDIST_{\mathrm{i},t}) + \delta_{\mathrm{i}} + \gamma_{\mathrm{t}} + \epsilon_{\mathrm{i,t}}$,

where $GROUP_i$ is a binary variable indicating whether a county belongs to a particular group of counties that may differ in the effects of railroads, and the other variables are as defined above. $Log(RAILDIST_{\mathrm{i},t})$ is instrumented as before by the log distance to the nearest connecting line, and the interaction term $GROUP_i * Log(RAILDIST_{\mathrm{i},t})$ is instrumented by the interaction between $GROUP_i$ and the log distance to the nearest connecting line.

Galor and Mountford (2008) provide an interesting hypothesis that can be tested using our natural experiment and this econometric model. According to the hypothesis, increasing trade might lead to different gains in different regions, due to specialization in different industries according to initial advantages. In regions that have a relative advantage in skilled-intensive industries trade will induce further investment in human capital and reduction of fertility rates, while in regions that have a relative advantage in unskilled-intensive industries the gains from trade might increase fertility rates.

Some of the counties in our sample, mostly in the Northeast, where relatively developed in 1850. Figure 12 presents a map of the counties that were relatively developed and relatively underdeveloped in 1850, according to their level of manufacturing output value per capita. Galor and Mountford hypothise that the

Figure 12: Initial Development Level of Counties



Notes: The figure shows the geographical distribution of counties in respect to the level of manufacturing value per capita in 1850. Other economic development variables provide similar results.

increase in trade induced by the railroads will have a larger effect on fertility and human capital in those counties. Table 11 establishes that this is indeed the case.

The table presents results for heterogeneity according to the initial conditions of economic development in 1850, as captured by 6 different groups of counties that represent 3 different definitions of development. The first 3 groups include counties that were relatively developed in 1850, while the last 3 groups include counties that were relatively underdeveloped in 1850. Groups 1 and 4, which appear in columns 1 and 4, were constructed by calculating the relative share of the value of manufacturing output per capita out of the sum of the value of manufacturing output per capita and the value of agricultural output per capita in 1850. The first group include the top 25% counties in respect to the share of manufacturing output per capita, and the fourth group include the bottom 25% counties in respect to this measure of industrialization. The groups in columns

2 and 5 include the top 25% counties in respect to the share of non-agriculture male workers in 1850 and the bottom 25% counties in respect to this measure. The groups in columns 3 and 6 include counties in the Northeast and counties in the South (no consistent results were found for the Midwest, and the West is not included in the sample group). The results are similar for other possible definitions of the groups.

Panel A in the table presents heterogeneity analysis for survival fertility and literacy. The results indicate that the effects were not just significantly larger (relative to all other counties) in counties that were relatively more developed in 1850, but also significantly smaller in the less developed counties. For example, while the effects of railroads on survival fertility and literacy in the Northeast is much larger than the general effect presented before, the effects in the South are not significantly different than zero. In column 2 the interaction coefficient is not significant for survival fertility and literacy, and for group 4 it is not significant for literacy, but the sign is the same as in the other interaction coefficients and the results are generally consistent.

Panels B and C further analyze the mechanisms behind this result, using the same 6 groups and different outcome variables. In panel B we analyze the hypothesis that the more initially developed counties specialized in manufacturing once they established access to railroads, while the underdeveloped counties specialized in agriculture. This hypothesis is not consistent with the results: most of the interaction coefficients are not significant, and some of the significant coefficients are in the opposite direction to what we might expect. For example, in Southern counties railroads increased farm value and agricultural output less than in other counties.

In panel C we analyze a broader hypothesis, according to which the more initially developed counties specialized in skilled-intensive industries once they established access to railroads, while the initially underdeveloped counties specialized in unskilled-intensive industries, which may or may not include

Table 11: Heterogeneity by Initial Level of Development

Group of Counties	(1) Top 25% manufacturing output share in 1850	(2) Top 25% non-agriculture workers share in 1850	(3) Northeast	(4) Bottom 25% manufacturing output share in 1850	(5) Bottom 25% non-agriculture workers share in 1850	(6) South
Panel A: main Outcomes						
Survival Fertility						
Distance to Railways	0.179***	0.184***	0.203***	0.188***	0.266***	0.0953***
	(0.0258)	(0.0318)	(0.0352)	(0.0252)	(0.0509)	(0.0168)
Group X Distance to Railways	0.0742***	0.0323	0.254**	-0.0594***	-0.0966***	-0.0860***
	(0.0266)	(0.0408)	(0.105)	(0.0132)	(0.0273)	(0.0102)
Literacy						
Distance to Railways	-0.119***	-0.123***	-0.131***	-0.120***	-0.154***	-0.0588***
	(0.0185)	(0.0256)	(0.0239)	(0.0179)	(0.0330)	(0.0216)
Group X Distance to Railways	-0.0425**	-0.0197	-0.167**	0.0160	0.0383**	0.0611***
	(0.0182)	(0.0356)	(0.0762)	(0.00978)	(0.0170)	(0.0150)
Panel B: Manufcaturing vs. Agriculture						
Manufacturing workers						
Distance to Railways	-0.131*	-0.132	-0.131*	-0.117*	-0.164	-0.130
	(0.0713)	(0.0900)	(0.0791)	(0.0706)	(0.117)	(0.113)
Group X Distance to Railways	0.0690	0.00763	0.0890	-0.135**	0.0350	0.0127
	(0.0563)	(0.119)	(0.0990)	(0.0599)	(0.0635)	(0.0791)
Farm value						
Distance to Railways	-0.791***	-0.780***	-0.811***	-0.801***	-0.999***	-0.437***
	(0.100)	(0.122)	(0.116)	(0.101)	(0.185)	(0.0719)
Group X Distance to Railways	-0.0912	0.0217	-0.223	0.0679	0.229**	0.369***
	(0.105)	(0.173)	(0.250)	(0.0587)	(0.0995)	(0.0469)
Agriculture output						
Distance to Railways	-0.240***	-0.197***	-0.238***	-0.246***	-0.262***	-0.0123
	(0.0405)	(0.0522)	(0.0451)	(0.0418)	(0.0684)	(0.0533)
Group X Distance to Railways	0.00630	0.179*	0.0227	0.0245	0.0239	0.257***
	(0.0522)	(0.102)	(0.0993)	(0.0252)	(0.0381)	(0.0389)
Panel C: Skilled-Intensive and Unskilled-Intensive Industries						
Share of workers in top 25% skilled-intensive industries						
Distance to Railways	-0.277***	-0.303***	-0.312***	-0.273***	-0.313***	-0.0818
	(0.0454)	(0.0630)	(0.0564)	(0.0438)	(0.0714)	(0.0619)
Group X Distance to Railways	-0.0973**	-0.0994	-0.513***	0.00338	0.0444	0.212***
	(0.0410)	(0.0858)	(0.166)	(0.0277)	(0.0365)	(0.0437)
Share of workers in bottom 66% skilled-intensive industries						
Distance to Railways	0.0856***	0.158***	0.103***	0.0872***	0.146***	0.000863
	(0.0228)	(0.0493)	(0.0281)	(0.0202)	(0.0387)	(0.0198)
Group X Distance to Railways	0.158***	0.258***	0.317***	-0.0452***	-0.0730***	-0.0864***
	(0.0339)	(0.0876)	(0.110)	(0.00899)	(0.0202)	(0.0118)
Share of workers in top 33% skilled-intensive industries (not including agriculture and missing industry)						
Distance to Railways	-0.107***	-0.129***	-0.119***	-0.112***	-0.156***	-0.0167
	(0.0235)	(0.0341)	(0.0280)	(0.0232)	(0.0372)	(0.0274)
Group X Distance to Railways	-0.0700***	-0.0801	-0.194***	0.0460***	0.0569***	0.0948***
	(0.0219)	(0.0489)	(0.0749)	(0.0126)	(0.0187)	(0.0177)
Share of workers in bottom 30% skilled-intensive industries (not including agriculture and missing industry)						
Distance to Railways	0.119***	0.132***	0.136***	0.124***	0.158***	0.0564**
	(0.0238)	(0.0328)	(0.0290)	(0.0224)	(0.0366)	(0.0272)
Group X Distance to Railways	0.0634***	0.0501	0.252***	-0.0425***	-0.0447**	-0.0666***
	(0.0212)	(0.0448)	(0.0826)	(0.0143)	(0.0186)	(0.0183)
County Fixed Effects	yes	yes	yes	yes	yes	yes
Year Fixed Effects	yes	yes	yes	yes	yes	yes

Notes: All variables are in logarithm except the dummies. Standard errors are clustered at the county level. Skilled and unskilled-intensive industries is measured according to the mean socioeconomic index per industry. The stars represent significance: *** p<0.01, ** p<0.05, * p<0.1.

agriculture. In order to analyze this hypothesis, 135 industries that appear in the 1850 full count data were divided into three groups, according to the average socioeconomic occupation index for males aged 25-65 who worked in those industries (which is also highly correlated with the literacy rate in each industry). The outcome variables are the share of workers in the top group and in the bottom group, for two different definitions of the groups: one that includes agriculture and "missing industry" (two large categories with relatively low skilled workers) and one that does not include those categories.

The first outcome in panel C is the share of workers in industries that employed 25% of the adult males in 1850 and had the highest average occupation index. This group include services industries such as legal services, health services and food stores, as well as manufacturing industries such as the printing and publishing industry and the leather industry. The second outcome in panel C include industries that employed 66% of the males in 1850 and had an average occupation index that is equal or lower than the index of the agriculture (we could not select the bottom group to be the same share as the top group because about 45% of the males were working in agriculture). 69% of the males in this group are working in agriculture, and other large industries include manufacturing industries such as yarn, thread, and fabric mills, mining and sawmills, and services industries such as shoe repair shops and taxicab services. The third outcome in panel C include the top industries that employed 33% of the workers in 1850 that didn't work in agriculture or had a missing industry, and the fourth outcome is the bottom industries that employed 30% of the workers in 1850 that didn't work in agriculture or had a missing industry.

According to the results presented in panel C, the interaction coefficients in the top groups are negative for top industries and positive for bottom industries, while the opposite is true for the bottom groups. For example, let's look at the top 25% skilled-intensive industries. Decreasing the distance to railroad by 10% increases the share of workers in those industries in the Northeast by 5.1% more

than in all other counties, and increases the share of workers in those industries in the South by 2.1% less than in all other counties. Looking at the bottom 66% industries, decreasing the distance to railroad by 10% decreases the share of workers in those industries in the Northeast by 3.17% more than all other counties, and decreases the share of workers in those industries in the South by 0.86% less than all other counties. The same pattern exists for other groups.

These results provide support for Galor and Mountford (2008). It seems that the mechanism is not about specialization in agriculture or in manufacturing but rather involves many services and manufacturing industries which are skilled-intensive or unskilled-intensive. These findings indicate on the existence of a positive feedback loop: in more developed counties railroads had a larger effect on fertility and human capital, which in turn may have encouraged further development, leading to a divergence between developed and less developed regions. This "small" divergence, which is a familiar phenomenon at the country-level during the Industrial Revolution (the "Great Divergence"), may not be visible in a "naïve" analysis due to confounding factors, unless we use a natural experiment.

6.2. Heterogeneity by Geographic Characteristics

Exogenous geographic characteristics might change the way railroads affect different regions. For example, railroads might have a smaller effect on individuals who live close to navigable rivers, so they were already connected to the national trade network before the arrival of the train.

Table 12 confirms that such patterns exist in the data: the effects of railroads on survival fertility and literacy are smaller for counties crossed by navigable rivers. The direction of the coefficients is the same for coastal counties, however the coefficient of the interaction is not significantly different than zero in this case. These patterns confirm that our identification strategy identifies the level of connectivity to the national trade network, and not something else.

Table 12: Heterogeneity by Geographic Characteristics

Group of Counties	(1) River Counties	(2) Coastal Counties
Panel A: Outcome = Survival Fertility		
Distance to Railways	0.192***	0.173***
	(0.0301)	(0.0290)
Group X Distance to Railways	-0.0310**	-0.00857
	(0.0135)	(0.0379)
Panel B: Outcome = Literacy		
Distance to Railways	-0.137***	-0.111***
	(0.0223)	(0.0239)
Group X Distance to Railways	0.0340***	0.0199
	(0.0104)	(0.0352)
County Fixed Effects	yes	yes
Year Fixed Effects	yes	yes

Notes: All variables are in logarithm except the dummies. Standard errors are clustered at the county level. The stars represent significance: *** p<0.01, ** p<0.05, * p<0.1.

Table 13: Heterogeneity by Initial Levels of Outcomes

Group of Counties	(1) Bottom 25% Survival Fertility in 1850	(2) Top 25% Literacy in 1850	(3) Top 25% Survival Fertility in 1850	(4) Bottom 25% Literacy in 1850
Panel A: Outcome = Survival Fertility				
Distance to Railways	0.226***	0.167***	0.179***	0.184***
	(0.0467)	(0.0221)	(0.0336)	(0.0240)
Group X Distance to Railways	-0.0502**	0.0348*	0.0137	-0.0535***
	(0.0242)	(0.0202)	(0.0457)	(0.0133)
Panel B: Outcome = Literacy				
Distance to Railways	-0.152***	-0.122***	-0.131***	-0.114***
	(0.0328)	(0.0161)	(0.0313)	(0.0180)
Group X Distance to Railways	0.0342**	0.0188	-0.0390	-0.0231**
	(0.0162)	(0.0119)	(0.0448)	(0.0107)
County Fixed Effects	yes	yes	yes	yes
Year Fixed Effects	yes	yes	yes	yes

Notes: All variables are in logarithm except the dummies. Standard errors are clustered at the county level. The stars represent significance: *** p<0.01, ** p<0.05, * p<0.1.

They also imply that the analysis presented in this paper cannot be done for later periods, when the popularity of automobiles increased, and highways became more important than railroads.

Other geographic characteristics, such as mean annual temperature, mean annual precipitation, mean altitude and land suitability, did not produce any consistent results.

6.3. Heterogeneity by Initial Level of Outcome Variables

Many counties had high rates of literacy already in 1850, and we would expect that the effect of railroads on literacy will be smaller in those counties. A similar argument might be made for the effect on fertility in counties where the fertility rate was already relatively low in 1850.

Table 13 confirms these hypotheses. According to column 1 of panel A the effect on survival fertility was smaller in low-fertility counties, and according to column 3 the effect was larger in the high-fertility counties. However, the coefficient of the interaction in column 3 is not statistically significant. Panel B presents similar patterns: the effects on literacy were larger in the less literate counties, and smaller in top literate counties. It is also interesting to note that the effect on fertility was larger in more literate counties and smaller in the less literate counties. This result is in line with the other results presented above for heterogeneity in respect to the initial development level.

The results of this analysis reinforce the results presented above for heterogeneity by initial development level: the effects of railroads in the most developed counties were larger even though human capital levels in those counties where already high to begin with, and fertility was already low to begin with.

7. Mechanisms

7.1. Immigration

A common problem in the literature based on county-level analysis is the effect of migrants on the results. The trends we document in fertility and literacy might be a result of a change in the behavior of local population in each county, or a result of selective migration from other counties. Migration might be especially problematic for our results if it changes the sex ratio, thus affecting fertility in a "mechanical" way, which is unrelated to the theories of the Demographic Transition we are interested in.

Columns 1-4 of Table 14 analyze the effect of railroads on outcomes that are relevant to immigration: the share of foreign born, the share of internal migrants, the sex ratio and population density. Unfortunately, data on migration between counties inside the same state is unavailable, so the share of internal migrants is calculated as the share of individuals in each county that were born in another state. While the share of foreign immigrants and the population density are negatively affected by the distance to the nearest railroad, the sex ratio is not affected, and the share of internal migrants actually decreases when we get closer to railroads. Thus, it seems that the results might have been partly driven by foreign immigrants, but not by the general patterns of interstate migration inside the US, and not through the sex-ratio mechanism.

Foreign born individuals during this period were more educated than natives and had a higher socioeconomic occupation index on average, but many of them settled in the major cities, and in most counties their share was small. Panel B of the table presents the distribution of the variables and analyzes the size of the effect, using a methodology similar to the one we used for the main results. As can be seen in column 1, the coefficient for the share of foreign born represent a relatively small effect. If we take a county that is in the bottom 25% in regards

Table 14: Mechanisms

	(1) IV	(2) IV	(3) IV	(4) IV	(5) IV	(6) IV
	Foreign Born	Internal Migrants	Sex Ratio	Population	Share of Young Married	Share of Females Living with Their Parents
Panel A: IV estimation						
Distance to Railways	-0.629***	0.190***	-0.00274	-0.340***	0.348***	-0.167***
	(0.0881)	(0.0401)	(0.00446)	(0.0505)	(0.0429)	(0.0244)
Distance to Cities	-0.331***	0.389***	-0.0437***	-0.401***	0.275***	-0.311***
	(0.0764)	(0.0442)	(0.00862)	(0.0565)	(0.0436)	(0.0309)
First stage F	60.90	65.79	61.46	61.46	62.01	67.87
County Fixed Effects	yes	yes	yes	yes	yes	yes
Year Fixed Effects	yes	yes	yes	yes	yes	yes
Panel B: Size of the effect						
Geographic variation in 1880:						
Bottom 25% value in 1880	0.3%	-	-	11,508	54.3%	30.6%
Median value in 1880	2.0%	-	-	17,950	46.4%	40.2%
Top 25% value in 1880	8.5%	-	-	28,124	37.3%	50.0%
Value if we start at the bottom 25% and decrease distance by 100%	0.6%	-	-	15,421	35.4%	35.7%

Notes: All variables are in logarithm except the dummies. Standard errors are clustered at the county level. The stars represent significance: *** p<0.01, ** p<0.05, * p<0.1.

to the share of foreign born and decrease the distance to zero, we will still be pretty far from the median. Furthermore, in the robustness tests we have seen that controlling for the share of foreign born does not alter the results of the estimation, so it seems the foreign immigrants were not an important part of the story.

The effect on population is also relatively small: a dramatic 100% decrease in the distance will increase the population size only by about a third, and will not get us to the median if we start at the bottom 25%. However, the decline of fertility rates might also affect the population size. Overall, it seems likely that a change in the behavior of the local population is a main driver for our results.

7.2. The Age of Marriage

One possible mechanism for the decrease in fertility is an increase in the age of marriage for males and females. Columns 5 and 6 of Table 14 presents results for two outcomes: the share of females aged 16-25 who are married, and the

share of females aged 18-25 that are still living with their parents. Both columns imply that the distance to railroads is negatively correlated with the age of marriage. Using males instead females produces similar results. The size of the effect on the share of young married is large: reducing the distance by 100% gets us from the bottom 25% (the larger share) to the top 25% group. The size of the effect on the share of females living with their parents is smaller.

This is not a "complete mechanism", in the sense that we don't know why males and females chose to marry at later ages. Males who worked as apprentices usually married only after finishing their training (in some cases marriage was forbidden by their contract), so one reason could be an increasing investment of males in their own human capital before marriage. Another possible explanation is an increase in the opportunity cost for marrying at young age for females, due to the new opportunities in the labor markets for unmarried females. Males and females might have also wanted to invest in their future children human capital, and chose to marry only when they accumulated enough wealth for that.

7.3. Female Labor

Galor and Weil (1996) argue that the decrease in the gender income gap during the 19th century could explain the Demographic Transition. The results presented in previous sections could reflect this mechanism, because many of the high ranked occupations employed relatively high shares of females. Unfortunately, data on female's occupations and female labor does not exist for the earlier years of our sample. Using the available data, no significant results were found regarding an effect on the literacy gender gap, or on some measures of female labor force participation in manufacturing. However, the lack of significant results might reflect missing data and not the lack of an effect in this case.

8. Concluding Remarks

This study shows that connecting 19th century Americans to the national trade network changed them: new opportunities in the labor market encouraged them to learn to read and write, to have fewer children and to marry at older ages. These results imply that railroads had an impact on the Demographic Transition, a dramatic change in human behavior that played a crucial role in the transition from the Malthusian stagnation to the modern era of constant economic growth (Galor, 2011).

The results are based on a novel identification strategy, that uses the emergence of new major cities to construct a dynamic instrument for the access to railroads, between 1850 and 1910. Combining the instrument with fixed effects for years and counties, I show that railroads led to a major shift in the distribution of occupations and industries, increased literacy rates, and had a negative effect on two different measures of fertility. The estimates of the effect are robust to different outcome variables, to different specifications of the econometric model, to different samples, and to different specifications of the instrument. Furthermore, they are not a result of pre-treatment trends that might violate the assumptions behind the identification strategy.

Heterogeneity analysis establishes that the effects of railroads were not homogenous. The increasing openness to trade induced specialization according to initial relative advantages: initially developed counties specialized in skilled-intensive industries, while initially underdeveloped counties specialized in unskilled-intensive industries - in line with the theory presented by Galor and Mountford (2008). Because of that, in counties that were initially more developed railroads had a relatively larger effect on fertility and human capital, while the opposite is true for counties that were initially less developed. In the long run, this mechanism might lead to a divergence between developed and less developed regions, similar to the Great Divergence between countries.

The arrival of railroads was accompanied by an increase in population density and in the share of foreign immigrants. However, the small size of those effects and a lack of effect on internal immigrants suggests that the effects were mainly driven by a change in the behavior of the local population. The effect of railroads on fertility was not a result of a change in the sex ratio, but it might have been driven by an increase in the age of marriage that accompanied the arrival of railroads.

While other studies also analyzed the effect of economic development on fertility and human capital (Pleijt, Nuvolari, and Weisdorf, 2016; Franck and Galor, 2017) or the effects of railroads in the US (Atack, Haines and Margo ,2008; Atack, Bateman, Haines and Margo, 2010; Donaldson and Hornbeck, 2016), this study adds to the literature by providing a novel identification strategy, and by creating a link between both strands of the literature - the impact of railroads and long-term growth.

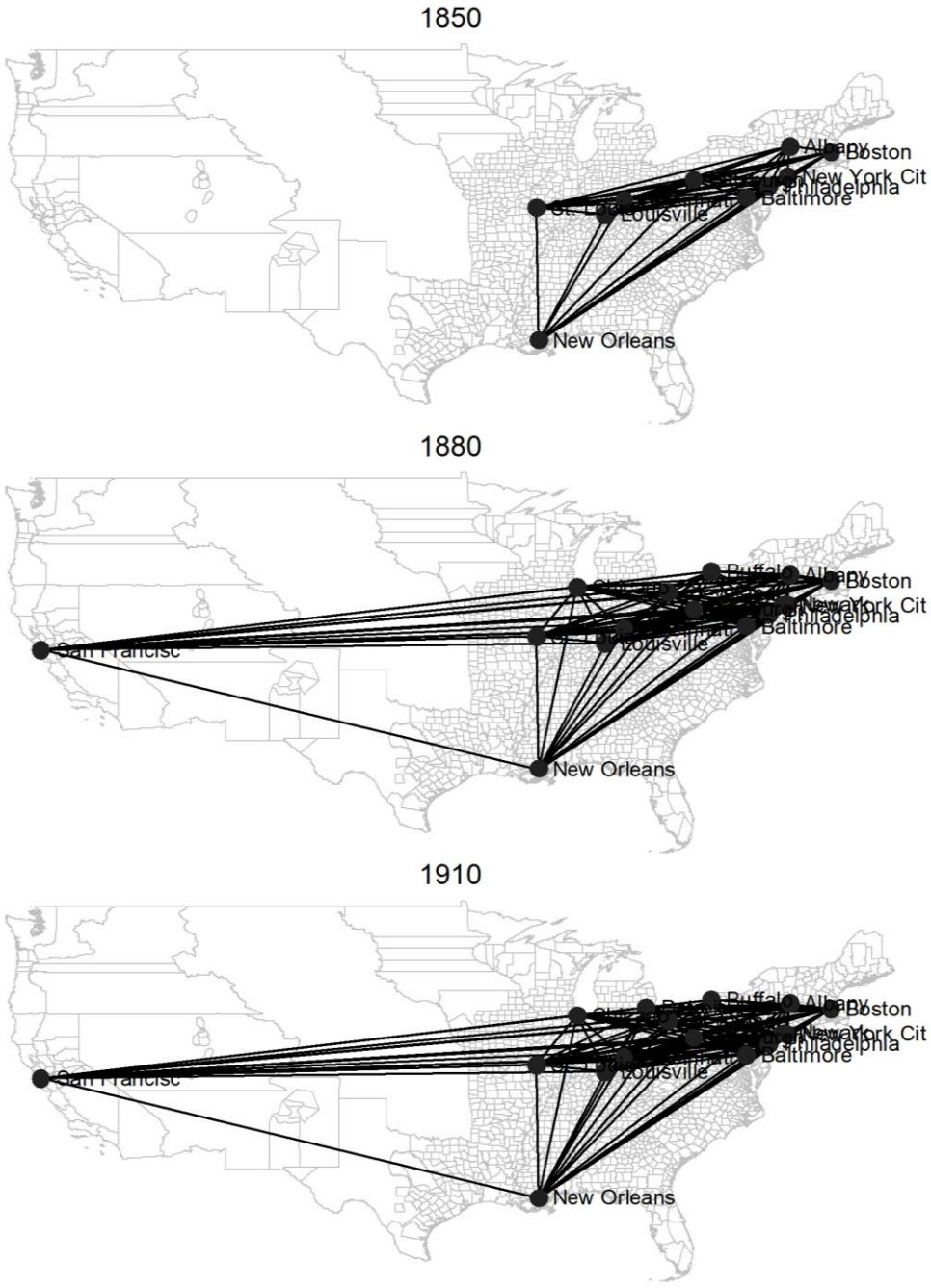
The results are generally in line with the literature. However, I find that the effect of railroads on the manufacturing sector was small and non-robust, compared to the general effects on industries and occupations or to what other studies find. It seems that the non-manufacturing sectors played an important role in raising the demand for human capital and driving the Demographic Transition, and the focus of the literature on the manufacturing sector might be misleading.

Today, globalization effects developing countries in ways that are similar to the effects of railroads on remote US counties. Studying the past allows us to have a better understanding of current trends, of the mechanisms that drive them, and of the possibility of heterogeneity in the effect of globalization.

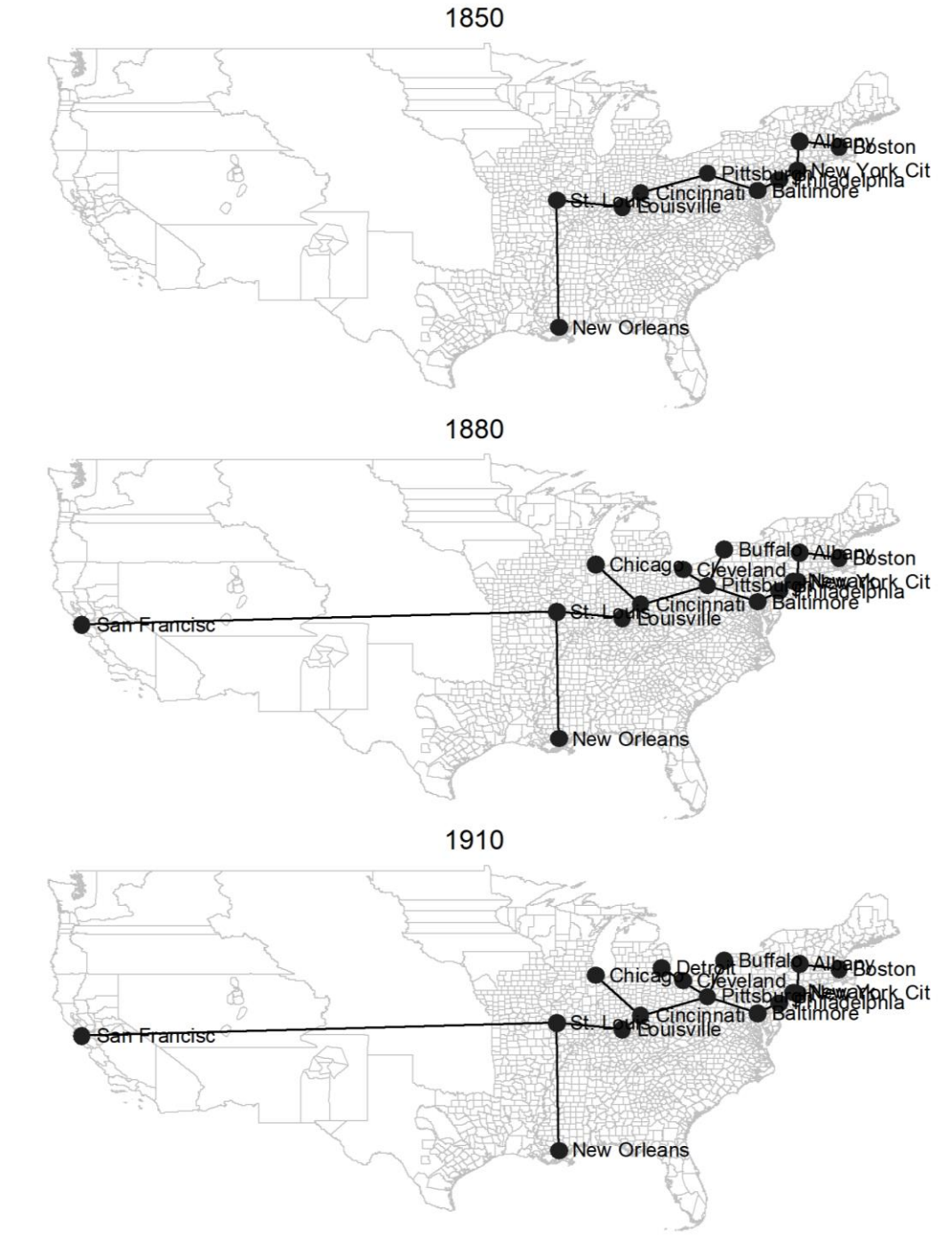
Online Appendix A: Connecting Lines Maps

Figure A1: Connecting lines in 1850, 1880 and 1910

Panel A: All Connecting Lines, 10 Major Cities



Panel B: MST Connecting Lines, 10 Major Cities



References

Atack, J., Bateman, F., Haines, M., & Margo, R. A. (2010). Did railroads induce or follow economic growth?. *Social Science History*, 34(02), 171-197.

Atack, J., Haines, M. R., & Margo, R. A. (2008). Railroads and the Rise of the Factory: Evidence for the United States, 1850-70 (No. w14410). National Bureau of Economic Research.

Banerjee, A., Duflo, E., & Qian, N. (2012). On the road: Access to transportation infrastructure and economic growth in China (No. w17897). National Bureau of Economic Research.

Becker, G. S. (1960). An economic analysis of fertility. In Demographic and economic change in developed countries (pp. 209-240). *Columbia University Press.*

Becker, G. S., & Lewis, H. G. (1974). Interaction between quantity and quality of children. In Economics of the family: Marriage, children, and human capital (pp. 81-90). *University of Chicago Press.*

Becker, S. O., Cinnirella, F., & Woessmann, L. (2010). The trade-off between fertility and education: evidence from before the demographic transition. *Journal of Economic Growth*, 15(3), 177-204.

Berger, T., & Enflo, K. (2017). Locomotives of local growth: The short-and long-term impact of railroads in Sweden. *Journal of Urban Economics*, 98, 124-138.

Cowan, R. S. (1997). A social history of American technology. *OUP Catalogue*.

Donaldson, D. (2018). Railroads of the Raj: Estimating the impact of transportation infrastructure. *American Economic Review*, 108(4-5), 899-934.

Donaldson, D., & Hornbeck, R. (2016). Railroads and American economic growth: A "market access" approach. *The Quarterly Journal of Economics*, 131(2), 799-858.

Duncan, O. D. (1961). A socioeconomic index for all occupations. Class: Critical Concepts, 1, 388-426.

Feldman, N. E., & van der Beek, K. (2016). Skill choice and skill complementarity in eighteenth century England. *Explorations in Economic History*, 59, 94-113.

Fernández, R. (2014). Women's rights and development. *Journal of Economic Growth*, *19*(1), 37-80.

Fishlow, A. (1965). American Railroads and the Transformation of the Ante-bellum Economy (Vol. 127). *Cambridge, MA: Harvard University Press*.

Fishlow, A. (1966). Levels of nineteenth-century American investment in education. *The Journal of Economic History*, *26*(04), 418-436.

Franck, R., & Galor, O. (2015). Industrialization and the Fertility Decline. *Brown University.*

Franck, R., & Galor, O. (2017). Technology-skill complementarity in early phases of Industrialization (No. w23197). National Bureau of Economic Research.

Galor, O. (2011). Unified growth theory. *Princeton University Press*.

Galor, O. (2012). The demographic transition: causes and consequences. *Cliometrica*, 6(1), 1-28.

Galor, O., & Moav, O. (2002). Natural selection and the origin of economic growth. *Quarterly Journal of Economics*, 1133-1191.

Galor, O., & Mountford, A. (2008). Trading population for productivity: theory and evidence. *The Review of economic studies*, 75(4), 1143-1179.

Galor, O., & Weil, D. N. (1996). The Gender Gap, Fertility, and Growth. *The American Economic Review*, 374-387.

Galor, O., & Weil, D. N. (1999). From Malthusian stagnation to modern growth. *The American Economic Review*, 89(2), 150-154.

Haines, M. R. (1998): "Estimated Life Table for the United States, 1850–1910," *Historical Methods, 31, 149–167.*

Hazan, M. (2009). Longevity and lifetime labor supply: Evidence and implications. *Econometrica,* 77(6), 1829-1863.

Hobsbawm, E. (2010). Age of Empire: 1875-1914. *Hachette UK*.

Hornung, E. (2015). Railroads and growth in Prussia. *Journal of the European Economic Association*, 13(4), 699-736.

Katz, L. F., & Margo, R. A. (2013). Technical change and the relative demand for skilled labor: The united states in historical perspective. In Human Capital in History: The American Record (pp. 15-57). *University of Chicago Press*.

Kennedy, P. (2010). The rise and fall of the great powers. *Vintage*.

Klemp, M. P., & Weisdorf, J. L. (2010). The child quantity-quality trade-off: evidence from the population history of England. *University of Copenhagen, mimeo.*

Landes, D. S. (2003). The unbound Prometheus: technological change and industrial development in Western Europe from 1750 to the present. *Cambridge University Press.*

Murphy, T. E. (2010). Old Habits Die Hard (Sometimes). Can département heterogeneity tell us something about the French fertility decline. *Bocconi University Innocenzo Gasparini Institute for Economic Research Working Paper, 364.*

Pleijt, A. M. D., Nuvolari, A., & Weisdorf, J. (2016). Human Capital Formation during the First Industrial Revolution: Evidence from the Use of Steam Engines (No. 294). *Competitive Advantage in the Global Economy (CAGE).*

Rosenberg, N., & Trajtenberg, M. (2004). A general-purpose technology at work: The Corliss steam engine in the late-nineteenth-century United States. *The Journal of Economic History*, *64*(01), 61-99.

Taylor, G. R. (1951). The transportation revolution, 1815-60. *Routledge.*

Wanamaker, M. H. (2012). Industrialization and fertility in the nineteenth century: Evidence from South Carolina. *The Journal of Economic History*, 72(1), 168-196.